Aus der Abteilung für Experimentelle Neurologie der Medizinischen Fakultät der Charité – Universitätsmedizin Berlin

Dissertation

Serielle Analyse der Genexpression (SAGE):

Etablierung, Statistik und Reliabilität

Zur Erlangung des akademischen Grades Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät Charité der Humboldt-Universität zu Berlin

Stefanie Castell aus München

Dekan: Prof. Dr. Martin Paul

Gutachter: 1. Prof. Dr. med. U. Dirnagl

2. Prof. Dr. rer. nat. Ch. Behl

3. Prof. Dr. K.- D. Wernecke

eingereicht: 28. Juli 2004

Datum der Promotion: 20. März 2006

Zusammenfassung

Die vorliegende Arbeit ist im Rahmen eines Projektes zur Untersuchung der Genexpression bei Tiermodellen neurologischer Erkrankungen entstanden. Mit herkömmlichen Kandidatenansätzen ist eine Genexpressionsanalyse nur in beschränktem Umfang zu realisieren. Ziel war daher die Etablierung eines Verfahrens wie SAGE (serielle Analyse der Genexpression), das die Analyse des gesamten Transkriptoms zuläßt. Wie die Arbeit zeigt, ist SAGE in einem Standardlabor durchführbar. Es wurden geringfügige Abwandlungen der Orginalmethode eingeführt. Zur Sequenzfehlerkorrektur wurde ein spezielles Computerprogramm entwickelt und evaluiert.
Zur Evaluierung der statistischen Auswertung von SAGE wurde zusätzlich zu einer Darstellung des gesamten statistischen Entscheidungsprozesses explorativ die Situation statistischer Entscheidungen wie sie im Rahmen üblicher SAGE Experimente auftreten mit vier Tests nachgeahmt. Es wurde eine Testvariante (modifizierter Z-Test) angewandt und evaluiert, die bis dato noch nicht zur Auswertung von SAGE benutzt worden war.
Um die Reliabilität von SAGE abschätzen zu können, wurde von vier Mäusegroßhirnen die Gesamt-RNS vereinigt. Diese Transkriptgrundpopulation wurde zweigeteilt und parallel untersucht. Die beiden Gruppen wurden anhand eines statistischen Tests, der die gesamte Verteilungen der beiden Profile prüft, auf Homogenität untersucht. Zusätzlich wurde das Zusammenhangsmaß ermittelt. Dies ergab, daß die Reliabilität von SAGE im vorliegenden Kontext (relativ geringe Stichprobe und ein komplexes Gewebe) nicht optimal ist. Es kann jedoch keine Aussage dazu gemacht werden, ob dies der Methode selbst, das heißt ihrer molekularbiologischen Praxis und der Datenaufbereitung, anzulasten ist oder einer großen Stichprobenvariabilität. Dies bedeutet, daß in der vorliegenden Arbeit keine endgültige Aussage zur Reliabilität von SAGE möglich ist. Es werden Möglichkeiten dargestellt mit einer suboptimale Reliabilität im Rahmen von zukünftigen Projekten umzugehen.

Abstract

The work presented here evolved within the framework of a project that examines gene expression of neurological conditions in animal models. Using conventional methods (candidate genes study) gene expression analysis is limited. Hence, the aim was to establish a procedure like SAGE (serial analysis of gene expression) that allows for analysis of the entire transcriptome. As shown it is possible to perform SAGE in a standard laboratory. Minor changes to the original version were made. A special computer program was developed and evaluated to reduce sequencing errors.

In addition to a description of the entire statistical process, the statistics of SAGE were explored by simulating normal SAGE experiments, using 4 statistical tests. One test version (modified Z-test) that has not been used for statistical analysis of SAGE yet was applied and reviewed.
To assess the reliability of SAGE the total-RNA of 4 corteces of mice was extracted and combined. This basic transcript population was divided in two and the parts examined in parallel. Both groups were analysed using a statistical test that tests the entire distribution of both profiles for homogeneity. Additionally the correlation (and its degree) of the profiles was calculated. The result was that the reliability of SAGE is not optimal in the context of this work (relatively small sample and complex tissue). However, no conclusion can be drawn as to whether the method itself (biomolecular practice and data analysis) is responsible for this, or whether it is due to sample variability. This means that in the work presented here no final statement concerning the reliability of SAGE is possible. Possibilities are described to deal with the issue of suboptimal reliability within the framework of future projects.

Schlagwörter:

SAGE, Reliabilität, Genexpression, mRNS

Keywords:

SAGE, reliability, gene expression, mRNA

Inhaltsverzeichnis

Zusammenfassung **2**
Abstract **2**
Abkürzungsverzeichnis **8**
1 Einleitung **9**
1.1 Einführung: "Wanderer, kommst du nach 7q31.2" (FAZ, 7.3.00) oder "Wohin mit drei Milliarden Buchstaben?" (FAZ, 9.10.01). 9
1.2 Definitionen 9
1.3 Hintergründe und Potential der Transkriptomanalyse 10
1.4 Verhältnis Boten-RNS und molekularer Phänotyp 12
1.5 Methoden der Transkriptomanalyse 13
1.5.1 Klassische Technologien der Genexpressionsanalyse 13
1.5.2 Neue Verfahren zur Untersuchung der gesamten Expression 13
1.5.3 Wahl einer geeigneten Methode zur globalen Untersuchung der Genexpression 15
1.5.4 Serielle Analyse der Genexpression (SAGE) 16
1.6 Verfahrensetablierung und Testgütekriterien 17
1.6.1 Nebengütekriterien 18
1.6.2 Hauptgütekriterien 19
1.7 Reliabilität 20
1.7.1 Reliabilität und ihre verschiedenen Komponenten 20
1.7.2 Gründe für die Messung der Reliabilität 23
1.7.3 Die Reliabilität von SAGE in der Literatur 23
1.8 Problemstellung und Hypothese 24
2 Materialien **26**
2.1 Kits 26
2.2 Vektoren und Bakterien 26
2.3 Enzyme 26
2.4 Chemikalien und weitere Stoffe 26
2.5 Puffer, Lösungen und Medien 28
2.6 Nukleotide, Linker und Primer 30
2.7 Geräte und Hilfsmaterial 31
2.8 Computer und Software 32
3 Methoden **33**
3.1 RNase freies Arbeiten 33

3.2 Steriles Arbeiten 33
3.3 Phenol - Choroform - Extraktion 33
3.4 3. 4 Ethanolpräzipitation von Nukleinsäuren 33
3.5 Bestimmung von Nukleinsäurenkonzentrationen 34
3.5.1 Messung der optischen Dichte 34
3.5.2 Semiquantitative DNS Mengenbestimmung mit Ethidiumbromid 34
3.6 Gelelektrophorese 34
3.6.1 Agarosegelelektrophorese 34
3.6.2 Polyacrcylamidgelelektrophorese 35
3.7 Restriktionsenzymverdau 35
3.8 Begradigen von cDNS 36
3.9 Ligation 36
3.10 Präparation kompetenter Bakterien und Anlegen einer Bakterienstammsuspension 36
3.11 Transformation per Elektroporation 37
3.12 Gesamt-RNS Isolierung aus Gewebe 37
3.13 Boten-RNS - Präparation mittels *Message Maker*™ Kit 38
3.14 cDNS - Herstellung 38
3.15 Polymerasekettenreaktion (PCR) 39
3.15.1 Allgemeine Prinzipien 39
3.15.2 PCR - Programme 42
3.16 Fluoreszein - Markierung von DNS - Sonden 42
3.16.1 Markierung in der PCR 42
3.16.2 Kontrolle des Fluoreszein-Einbaus 43
3.17 Northern - Blot 43
3.17.1 Transfer der RNS auf eine Nylonmembran 43
3.17.2 Hybridisierung und Detektion von Northern-Blot 44
3.18 cDNS Southern - Blot 45
3.18.1 Transfer von cDNS auf eine Nylonmembran 45
3.18.2 Hybridisierung und Detektion von cDNS Southern -Blots 46
3.19 Präparative Gelelektrophorese 46
3.19.1 QIAqick Gel Extraction Kit 46
3.19.2 Präparation von DNS-Fragmenten nach Polyacrylamidgelelektrophorese 46
3.20 Binden von biotinylierten cDNS-Fragmenten an *Dynabeads* 47
3.21 Vektorisolation 47

3.21.1 Reinigen von Plasmid-DNS mit dem *Plasmid Mini Purifikation Kit* (Qiagen) 47
3.21.2 Reinigen von Plasmid-DNS mit *Qiaprep 96 Turbo Miniprep Kit* 48
3.22 Sequenzierung 48
3.22.1 Sequenzierung mit fluoreszierenden Primern 48
3.22.2 Sequenzierung mit fluoreszierenden Nukleotiden 49
3.23 Statistische Methoden 49
4 Etablierung von SAGE 50
4.1 Ergebnisse der Etablierung 50
4.1.1 Ausgangssituation und Ziel 50
4.1.2 Zusammenfassende Beschreibung der Methode 50
4.1.3 Vorbereitende Tests 52
4.1.4 SAGE-Durchlauf 55
4.2 Diskussion der Etablierung von SAGE 81
4.2.1 Methodische Probleme der Durchführung von SAGE 81
4.2.2 Probleme der Auswertung 88
4.2.3 Implikationen für SAGE bei Sonderfällen der Boten-RNS 101
4.2.4 Beurteilung des quantitativen Resultats 105
4.3 Fazit der Praxis von SAGE 108
5 Statistische Evaluation und Reliabilität von SAGE 112
5.1 Ergebnisse 112
5.1.1 Ausgangsituation und Strategie 112
5.1.2 Vergleich der Gesamtverteilungen 113
5.1.3 Paarweise Vergleiche 119
5.1.4 Zusammenfassung der Ergebnisse der Simulationen üblicher Experimente 126
5.1.5 Statistischer Vergleich der angewandten paarweisen Tests 127
5.1.6 Zusammenfassung der Berechnungen zur Reliabilität von SAGE 130
5.1.7 Sequenzfehlerkorrektur 131
5.2 Diskussion 133
5.2.1 Der statistische Entscheidungsprozeß 133
5.2.2 Evaluation nicht angewandter Tests 141
5.2.3 Evaluation der angewandten Tests 145
5.2.4 Reliabilität 155
6 Zusammenfassung 176
Literaturverzeichnis 180

Danksagung.....189
Lebenslauf.....189
Eidestattliche Erklärung.....189

Abkürzungsverzeichnis

A	Adenin
Acc. No.	(*accession number*) Nummer der Sequenzen in den Datenbanken
AKAP	Ankerproteine für die Proteinkinase A (*A kinase anchoring protein*)
bp	Basenpaare
BSA	Albumin aus Rinderserum (*bovine serum albumin*)
C	Cytosin
cDNS	aus Boten-RNS hergestellte DNS
dNTP	Deoxyribonukleotidtriphophat
DEPC	Diethylpyrokarbonat
DMSO	Dimethylsulfoxid
DNS	Desoxyribonukleinsäure
DNase	Desoxyribonuklease
EDTA	Ethylendiamintetraazetat
EST	nicht näher identifizierte Sequenzen (*expressed sequence tag*)
G	Guanin
GS	Größenstandard
Insert	in einen Vektor einkloniertes DNS-Fragment
LB	*Luria broth*
M	molar
PAGE	Polyacrylamidgelelektrophorese
PCR	Polymerasekettenreaktion (*polymerase chain reaction*)
RNS	Ribonukleinsäure
Rnase	Ribonuklease
RT	Raumtemperatur
SAGE	Serielle Analyse der Genexpression
SNP	Einzelnukleotidpolymorphismus (*single nucleotid polymorphism*)
TAE	Tris-, Essigsäure- und EDTA-haltiger Puffer
Tag	kurze Sequenz eines Transkriptes
Taq	DNS-Polymerase aus *Thermus aquinatus*
TE	Tris-EDTA-Puffer
TEMED	N, N, N', N' - Tetramethylethylendiamin
Tween 20	Polyoxyethylensorbitanmonolaurat

1 Einleitung

1.1 Einführung: "Wanderer, kommst du nach 7q31.2" (FAZ, 7.3.00) oder "Wohin mit drei Milliarden Buchstaben?" (FAZ, 9.10.01).

"Doch noch ein deutsches Genomprojekt. Aufholjagd beginnt mit der Entwicklung von Technologien. Im Juni des vergangenen Jahres wurde das deutsche Genomprojekt aus der Taufe gehoben." (FAZ, 21.2.96), "Richtfest: Die Entschlüsselung des Erbguts. 10. Januar 2000." (FAZ, 30.12.00), "Informationsflut aus dem Erbgut. Mehr als 135 Lebewesen vor der genetischen Entzifferung ." (FAZ, 19.4.00), "Goliath ganz groß. Das Genomprojekt floriert." (FAZ, 22.5.00), "Den Deutschen ist das Genom wichtiger als die Mondlandung." (FAZ, 17.8.00), "Wer soll das alles lesen? Noch ertrinken auch Spitzenforscher in den Buchstaben des Genoms, [...]" (FAZ, 10.2.01), "Entschlüsselung der Banane naht" (FAZ, 19.07.01), "Materialschlacht um die Bausteine des Lebens. Das "Humanproteom-Projekt" wird von Ingenieuren und Industrie immer schneller vorangetrieben." (FAZ, 13.2.02).

Diese kleine Chronologie von Schlagzeilen zur Genomanalyse deutet die Brisanz des Themas und das Tempo der Entwicklung auf diesem Gebiet an. Die Entschlüsselung des gesamten menschlichen Genoms und des Genoms anderer Organismen wie der Maus ist inzwischen abgeschlossen (Internationales Human Genome Sequencing Consortium 2001). Doch die eigentliche Arbeit beginnt damit erst. Denn um die komplexen physiologischen und krankhaften Vorgänge auf molekularer Ebene zu verstehen, reicht die vollständige Sequenzierung des Genoms nicht aus. Daher rückt in der gegenwärtigen Postgenomära (Bertelsen und Velculescu 1998) die Analyse der Funktionsweise des Genoms und damit die Untersuchung des sogenannten Transkriptoms in den Fokus medizinisch-biologischen Interesses. In diesen Kontext ist auch die vorliegende Arbeit einzuordnen, da sie eine der zentralen Methoden der Transkriptomanalyse - SAGE (serielle Analyse der Genexpression) - am Beispiel des murinen Gehirns untersucht und unter besonderer Berücksichtigung der Reliabilität evaluiert.

Um die Hintergründe und Potentiale der Postgenomära und von SAGE verständlich erläutern zu können, werden im folgenden die Definitionen einiger zentraler Begriffe geklärt.

1.2 Definitionen

Der Begriff Genom ist mehr als 75 Jahre alt und bezeichnet den kompletten Gen- und Chromosomensatz eines Organismus. Das Wort '*genomics*' (Genomanalyse) wird seit 1986

verwendet und beinhaltet das Kartieren, Sequenzieren und Analysieren eines Genoms. Im Laufe der 90er Jahre fiel unter diesen Begriff zunehmend auch die Untersuchung der Funktionen des Genoms. Inzwischen wird die strukturelle Genomanalyse von der funktionellen ("functional genomics") unterschieden. Während erstere die Kenntnisse der kompletten DNA Sequenz und (abnormer) Varianten des Genoms zum Ziel hat, widmet sich letztere der systematischen Untersuchung des Transkriptoms und Proteoms (Hieter und Boguski 1997, Fields et al. 1999) und damit dem Verstehen der Funktionsweise des Genoms. Das Transkriptom stellt die Gesamtheit der Boten-RNS dar, welche zu einem gegebenen Zeitpunkt in einer definierten Zellpopulation vorhanden ist. Es umfaßt alle Transkriptionsprodukte der aktiven Gene dieser Zellen (Velculescu et al. 1997). Durch die Translation der Boten-RNS an den Ribosomen entsteht aus dem Transkriptom das Proteom, welches wiederum sämtliche Proteine umfaßt.

1.3 Hintergründe und Potential der Transkriptomanalyse

Aus dreierlei Gründen ist es interessant, das Transkriptom in seiner Gesamtheit zu untersuchen. Zum einen können bisher unbekannte Gene entdeckt und ihre Funktionsbereiche eingegrenzt werden. Zweitens können durch die Bestimmung differentiell exprimierter Gene Interessensschwerpunkte herausgearbeitet werden. Forschung soll auf diese Weise gezielt durchgeführt werden können und damit beschleunigt werden. Drittens können durch die Untersuchung des gesamten Transkriptoms umfassende Erkenntnisse über physiologische und krankhafte Prozesse auf molekularer Ebene gewonnen und medizinisch genutzt werden. All diese Forschungsmotive sollen im Folgenden kurz erläutert werden.

Entdeckung bisher unbekannter Gene und Eingrenzung ihrer Funktion

Viele Gene, ihre korrespondierenden Boten-RNS-Varianten sowie ihre Funktionen sind noch nicht bekannt. Durch eine Analyse des gesamten Transkriptoms läßt sich diese Wissenslücke verringern. Veculescu et al. (1999) zeigten in einer Metaanalyse von 84 menschlichen Transkript-Bibliotheken, die mittels einer Methode zur Transkriptomanalyse (serielle Analyse der Genexpression, SAGE) erstellt wurden, daß von 84000 Transkriptvarianten 46% nicht in eine Gendatenbank eingetragen sind. Eine Untersuchung von fast 30000 Transkripten des menschlichen Gehirns (Lal et al. 1999) ergab, daß 23% dieser Boten-RNS Sequenzen keinem Gen in einer Gendatenbank zuzuordnen waren. Die SAGE Analyse des Transkriptoms von *Saccharomyces cervisicae* (Velculescu et al. 1997) ermöglichte die Identifizierung mehrerer hundert neuer Gene, welche sich zuvor bei der computergestützten Analyse des Genoms anhand

von offenen Leserahmen nicht hatten vorhersagen lassen. Der experimentelle Nachweis von Boten-RNS per SAGE macht die Annahme wahrscheinlich, daß sich die computergestützten Untersuchungen auf Genomebene zur Vorhersage von Genen nur begrenzt eignen und die Annahme von 30.000 Transkripten beziehungsweise Transkriptvarianten für das menschliche Genom verdoppelt werden muß (Chen et al. 2002). Diese Beispiele - eine Liste, die sich noch wesentlich verlängern ließe - zeigen, daß hier noch ein enormer Forschungsbedarf besteht.

Bestimmung primär interessanter Gene

Durch die Analyse der differentiellen Genexpression lassen sich aus der Masse bereits vorhandener Daten, Gene auswählen, die dadurch, daß sie in bestimmten Zusammenhängen signifikant reguliert erscheinen, besondere Wichtigkeit indizieren. Diesen Genen kann dann in weiteren Untersuchungen Priorität gewährt werden. Durch eine solche gezielte Vorgehensweise verspricht man sich eine Beschleunigung medizinisch-biologischer Forschung. Durch Datenbanken, die aus großangelegten Genexpressionsanalyseprojekten stammen und per Internet öffentlich zugänglich sind (Hough et al. 2000, Lal et al. 1999, Velculescu et al. 1999), wird diese Strategie zusätzlich unterstützt. So können Daten zur Genexpression von mehr WissenschaftlerInnen in Analysen miteinbezogen werden. Außerdem können durch die dort vorhandene Datenbreite Expressionsmuster in einem weitaus größeren Umfang herausgearbeitet werden, als das bisher möglich war.

Charakterisierung molekularer Mechanismen und deren medizinische Nutzung

Es wird davon ausgegangen, daß fast jeder biologische Prozeß mit Veränderungen in der Genexpression assoziiert ist (Velculescu et al. 1997). Durch die Analyse der differentiellen Expression können daher Einblicke in molekulare Abläufe bestimmter Krankheiten oder pathologischer Zustände gewonnen werden und diese charakterisiert werden. So konnten Polyak et al. (1997) auf diese Weise ein Modell für p53 induzierte Apoptose aufstellen.

Zusätzlich lassen sich potentielle Ansatzpunkte für therapeutische und diagnostische Strategien beleuchten. Für einen pulmonalen und einen pankreatischen Tumortypus wurden beispielsweise anhand der Analyse der entsprechenden Transkriptome mittels SAGE neue diagnostisch und prognostisch sinnvolle Markergene gefunden (Madden et al. 2000). Lal et al. (1999) weisen auf die Möglichkeit hin, durch Transkriptomanalyse Kandidatengene für Tumorimmuntherapien zu finden. Auch für die Gentherapie kann die umfassende Kenntnis der gewebsspezifischen Genexpression für das therapeutische Vorgehen entscheidend sein.

Angesichts der Tatsache, daß nur 4% der klassischen 'drug discovery' Projekte letztendlich neue Medikamente auf den Markt bringen (Madden et al. 2000), ist eine Entwicklung effektiverer neuer pharmakologischer Ansätze wünschenswert. Diese setzen jedoch im Gegensatz zur klassischen Vorgehensweise, die Substanzen nach möglichen therapeutischen Effekten auf viele verschiedene Angriffsstellen abtastet, eine bessere Kenntnis der molekularen (Patho-) Physiologie voraus. Anhand in diesem Bereich neu gewonnen Wissens kann dann die Forschung gezielt ihre Bemühungen auf biologisch relevante therapeutische Ansatzpunkte richten.
Des weiteren können Substanzen bereits während einer frühen Entwicklungsphase auf Einflüsse in der Genexpression getestet werden, was die Effizienz pharmakologischer Forschung weiter steigern und die Kosten sowie die Anzahl der Versuchstiere für die Entwicklung neuer Medikamente senken würde (Cox 2001, Madden et al. 2000).

1.4 Verhältnis Boten-RNS und molekularer Phänotyp

Angesichts all dieser Vorteile, die umfassende Untersuchungen der Genexpression wie oben erläutert in Aussicht stellen, gilt es sich dem Einwurf zu stellen, daß das Niveau der Boten-RNS nicht oder nicht exakt mit dem Niveau der entsprechenden Proteine und damit der Funktion einer Zelle einhergeht. Dies würde den Sinn aller aufwendigen Transkriptomuntersuchungen in Frage stellen und die dargestellten Beweggründe nichtig machen.
Der molekulare Phänotyp einer Zelle oder eines Zellverbandes wird nicht nur von der Quantität der Boten-RNS, sondern auch von deren Stabilität und derjenigen der korrespondierenden Proteine und von anderen post-transkriptionale oder post-translationalen Regulations-mechanismen beeinflußt. Die Kenntnis der Menge einer Boten-RNS erlaubt daher keine 100% präzise Vorhersage über die Menge des entsprechenden Proteins und über den molekularen Phänotyp (Gygi et al. 1999). Dennoch kann sie als Indikator für Zellvorgänge fungieren. Es müssen - bei Bedarf, das heißt, falls ein solcher Hinweis nicht ausreichend ist - weiterführende Studien folgen. Chen et al. (1998) zeigten beispielsweise mit einer Untersuchung (SAGE) zur allergischen Mastzellaktivierung, daß die Expression von MIF Boten-RNS (Makrophagen-Migrationsinhibitionsfaktor) der Menge des zugehörigen Proteins entspricht (Western Blot), was auf eine Beteiligung von Mastzellen an der Infektionsabwehr hinweist. Diese Funde wurden in Studien mit Mäusen, welchen es an Mastzellen mangelte, bestätigt.
Wenn man die komplexen (patho-)physiologischen Vorgänge auf molekularer Ebene genau verstehen möchte, ist die Kenntnis der Transkription (qualitativ und quantitativ) zudem wesentlich, weil sie einen essentiellen Schritt in der Achse Genom - molekularer Phänotyp

darstellt. Ohne die Kenntnis der Vorgänge auf der Ebene der Boten-RNS würde in der Erforschung zellulärer Ereignisse eine Wissenslücke existieren.

1.5 Methoden der Transkriptomanalyse

Es stellt sich nun die Frage, welche Methoden zur Verfügung stehen, um eine Transkriptomanalyse durchzuführen.

Angesichts der Komplexität der Genexpression - es werden abhängig vom Gewebe bis zu Zehntausende von Genen in je unterschiedlicher Menge abgeschrieben (Internationales Human Genome Sequencing Consortium 2001) - ist die Charakterisierung des Transkriptoms kein simples Unterfangen und erfordert Verfahren, welcher dieser Komplexität gerecht werden können. Im folgenden werden einige zentrale Methoden der Genexpressionsanalyse vorgestellt und bezüglich ihrer Tauglichkeit für die Untersuchung des gesamten Transkriptoms diskutiert, um so deutlich zu machen, weshalb für die vorliegende Untersuchung SAGE gewählt wurde.

1.5.1 Klassische Technologien der Genexpressionsanalyse

Klassische Verfahren wie zum Beispiel die seit 1977 (Kozian und Kirschbaum 1999) existierende Northern Blot Analyse beruhen auf Kandidatenansätzen. Das heißt, daß sie lediglich in Einzelexperimenten die Expression bereits bekannter Gene analysieren. Derartige Methoden sind bei weitem zu arbeits- und materialintensiv, um für die primäre Analyse des gesamten Transkriptoms eine Bedeutung zu haben. Ihre Stärke liegt in der sekundären Validierung von Daten, welche mit neuen komplexeren Methoden der pauschalen Transkriptomanalyse erhoben wurden.

1.5.2 Neue Verfahren zur Untersuchung der gesamten Expression einer Zellpopulation

In den letzten zehn bis fünfzehn Jahren wurden Verfahren entwickelt, die nicht die Expression einzelner Kandidatengene untersuchen, sondern versuchen, das Transkriptom in seiner Gesamtheit zu erfassen.

Diese Verfahren lassen sich in geschlossene und offene Systeme einteilen. Erstere müssen sich auf die Analyse bereits bekannter Gene beschränken, während sich die offenen Techniken dadurch auszeichnen, daß sie bisher unbekannte Gene in ihre Untersuchung mit einbeziehen können und keinerlei schon vorhandene Information über die Sequenz oder den biologischen

Hintergrund des Untersuchungsgegenstandes benötigen (Green 2001).
Des weiteren können diese neuen Verfahren anhand ihre Meßweise in zwei Gruppen geteilt werden. So werden analoge, das heißt ablesende Techniken, die ein figuratives Signal produzieren, das durch seine Intensität die Stärke der Genexpression widerspiegelt, von digitalen, das heißt zählenden Methoden unterschieden.
Im Folgenden werden häufig verwendete Verfahren zur Untersuchung der gesamten Genexpression einer Zellpopulation unter dem Gesichtspunkt ihrer Meßweise zusammengefaßt (siehe auch Green et al. 2001, Kozian und Kirschbaum 1999 und Carulli et al. 1998).

Analoge Verfahren

Methoden dieser Art sind in Tabelle 1 aufgeführt, wobei die Chip- und Filtertechniken bei weiten am häufigsten zur Untersuchung der Genexpression angewandt werden.

Methode	**System**	**Literatur**
Chip- und Filtertechniken	geschlossen	Überblick: Cox (2001), Ivanov et al. (2000), Lockhard und Winzler (2000)
Differential Display	offen	Liang und Pardee (1992), Jakob et al. (1999)
Subtraktives Klonieren	offen	Hubank und Schatz (1994)
TOGA	offen	Sutcliffe et al. (2000), Thomas et al. (2001)
GeneCalling	offen	Shimkets et al. (2001)

Tabelle 1. Analoge Methoden.

Bei den analogen Verfahren ergibt sich aus der Übersetzung der Intensität des figurativen Signals in Zahlenwerte ein wesentlicher Nachteil gegenüber direkt quantifizierenden, das heißt digitalen Methoden (Audic und Claverie 1997, Spanakis und Bouty - Boyé 1994, Spanakis 1993). Durch die Tatsache, daß die Quantifizierung der Genexpression erst sekundär erfolgt, besteht hier im Vergleich zu digitalen Verfahren eine weitaus größere Gefahr der Ungenauigkeit der quantitativen Resultate.
Um Vergleiche innerhalb eines Experimentes möglich zu machen, sind darüber hinaus diverse aufwendige Normalisierungs- und Standardisierungsprozesse notwendig, da das Niveau der Genexpression bei dieser Verfahrensweise nur relativ gemessen wird.

Digitale Verfahren

Die gebräuchlichsten digitalen Methoden faßt Tabelle 2 zusammen. Die Attraktivität dieser digitalen Verfahren liegt darin, daß sie im Gegensatz zu analogen Methoden die Genexpression unmittelbar zählen und so Boten-RNS Mengen absolut quantifizierbar machen. Die Gefahr unpräziser Meßergebnisse sollte also theoretisch bei einer solchen Vorgehensweise niedriger sein als bei analogen Verfahren. Wie aus der Tabelle 2 hervorgeht, bieten außerdem alle genannten Verfahren die Möglichkeit, neue Transkripte zu erfassen, das heißt, daß es sich um offene Systeme handelt.

Methode	**System**	**Literatur**
"expressed sequence tags" (EST) Sequenzierung	offen	Adams et al. (1991 und 1995), Wilcox et al. (1991)
SAGE	offen	Velculescu et al. (1995)
Oligonukleotid Fingerprinting	offen	Meier-Ewert et al. (1998)
MPSS	offen	Brenner et al. (2000)

Tabelle 2. Digitale Methoden.

1.5.3 Wahl einer geeigneten Methode zur globalen Untersuchung der Genexpression

Zur Umsetzung des Ziels, Transkriptome umfassend zu untersuchen, wurde die Etablierung folgender Methoden diskutiert: Filtertechniken, subtraktives Klonieren, Differential Display, EST Sequenzierung und SAGE.

Wie bereits erwähnt stellen Chip- und Filterverfahren häufig verwendete analog arbeitenden Techniken dar. Expressionsprofile können hier durch parallele Analyse vergleichsweise schnell für eine Vielzahl verschiedener Gewebe, Pathologien und experimenteller Bedingungen (Heller et al. 1997, Welsh et al. 2001) und - je nach Subtyp der Technologie - auch für eine großen Anzahl von Genen erfaßt werden. Deswegen wird dieser Methodik eine Zukunft als "zentrale Plattform" der funktionellen Genomanalyse vorausgesagt (Fields et al. 1999, Schena et al. 1998). Ohne im Besitz automatisierter Vorrichtungen zu eigener Herstellung zu sein, ist man jedoch auf die kommerziell vertriebenen Filter oder Chips angewiesen (zum Beispiel Atlas™ Array, Clontech). Dies bedeutet, auf die parallele Analyse einer begrenzten Anzahl von bereits

bekannten und vorausgewählten Genen beschränkt zu sein.

Zwei weitere analoge Verfahren - Differential Display und subtraktives Klonieren - eignen sich tendenziell nur zur Erforschung der qualitativen Seite der Genexpression (Madden et al. 1997). Zudem wurde von beiden Verfahren eine hohe Rate an fasch positiven Ergebnissen berichtet. Bei Differential Display kann diese bei bis zu 50% liegen (Martin und Pardee 2000). Zudem wird berichtet, daß diese Methode nicht sehr zuverlässig sei (Spinella et al. 1999). Subtraktives Klonieren ist außerdem durch das Problem, viele falsch negative Resultat zu liefern, gekennzeichnet (Hubank und Schatz 1999). Wenn man die Grenzen dieser analogen Verfahren nicht akzeptieren möchte, bleiben die beiden genannten digitalen Methoden (SAGE und EST Sequenzieren) zur Auswahl, wobei SAGE durch den höheren Durchlauf gegenüber dem reinen EST Sequenzieren wesentliche Vorteile bietet (Larsson et al. 2000, Man et al. 2000, Bertelsen und Velculescu 1998, Audic und Claverie 1997). Um dies zu verdeutlichen, werden im folgenden Abschnitt die grundlegenden Prinzipien von SAGE skizziert.

1.5.4 Serielle Analyse der Genexpression (SAGE)

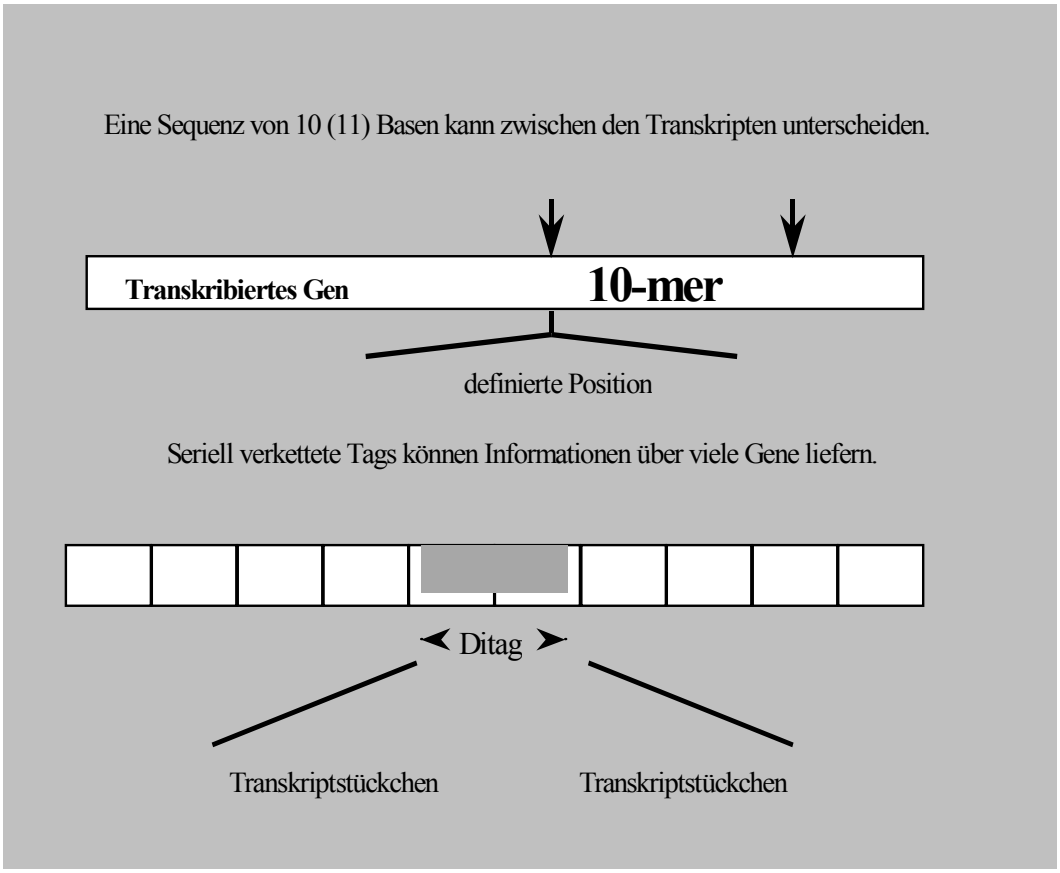

Abb. 1: Das Prinzip von SAGE.

Das Prinzip der seriellen Analyse der Genexpression basiert auf der Herstellung von 14 bis 15

Basen langen Fragmenten, sogenannten *'tags'* (siehe Abb. 1). Diese sind einer definierten Position am 3'- Ende der Boten-RNS entnommen und charakterisieren so spezifische Transkripte. Diese kurzen Fragmente werden zu langen Ketten ligiert, kloniert, seriell sequenziert und durch Auszählung quantifiziert. Durch einen Vergleich dieser kurzen Sequenzen mit in Genbanken vorhandenen Daten werden bereits bekannte Gene identifiziert. Variationen in der Häufigkeit der Tags zwischen verschiedenen Expressionsprofilen können dazu benutzt werden, um statistisch signifikante Unterschiede in der Genexpression zwischen beispielsweise gesunden und pathologisch verändertem Gewebe herauszuarbeiten. Gegenüber dem EST Sequenzieren ergibt sich aus der Kürze der Tags und deren Verkettung für SAGE der Vorteil, daß seriell - und je nach vorhandenen Kapazitäten auch parallel - sequenziert werden kann, was einen wesentlich höheren Durchlauf ermöglicht. So bietet bei SAGE die Sequenzierung eines Klons die Identifizierung von 30 - 50 Boten-RNS (Tyagi 2000), beim EST Sequenzieren jedoch nur von einer RNS. Dies ist für die Herausarbeitung von statistisch signifikanten Expressionsunterschieden essentiell und senkt die Kosten pro Gen (Man et al. 2000, Spinella et al. 1999). Die Möglichkeit, Transkripthäufigkeiten (an Stichproben) absolut zu zählen, läßt bei standardisiertem Datenmanagement Metaanalysen zu (Velculescu et al. 1999, Lal et al. 1999). Als Methode zur Analyse des Transkriptoms ist SAGE nicht nur dadurch attraktiv, daß dieses Verfahren die Genexpression umfassend und differenziert erfassen kann, sondern auch als offenes System bisher unbekannte Gene entdecken kann. SAGE wird aus all diesen Gründen als der beste Vertreter der offenen Systeme bezeichnet (Nacht et al.1999). Dies hat zur Folge, daß die Anwendung von SAGE weltweit zunimmt (Ye und Zhang 2000) und die Anzahl der per SAGE analysierten Transkripte im Jahr 1999 fast fünf Millionen betrug (Yamamoto et al. 2001).
In unserer Arbeitsgruppe sollte daher diese Methode etabliert werden mit dem Ziel, Fragen der Genexpression in relevanten Tier- und Zellkulturmodellen (zum Beispiel zerebralen Ischämie) umfassend beantworten zu können. Bevor jedoch derartige inhaltliche Fragestellungen angegangen werden können, sollte - wie bei allen neu entwickelten Verfahren - sichergestellt werden, daß SAGE wissenschaftlichen Gütekriterien entspricht. Im Folgenden sollen diese Kriterien kurz zusammengefaßt werden. Zugleich soll dargestellt werden, ob und inwiefern SAGE sie erfüllt, beziehungsweise ob hier noch Forschungsbedarf besteht.

1.6 Verfahrensetablierung und Testgütekriterien

Die Gütekriterien eines Verfahrens lassen sich in Haupt- und Nebenkriterien unterteilen. Bei ersteren handelt es sich um Objektivität, Reliabilität und Validität. Die Nebengütekriterien -

Utilität, Ökonomie und Vergleichbarkeit - sind vor allem bezüglich der praktischen Umsetzbarkeit des Verfahrens von Bedeutung (Weineck 1996[9]). Als Verfahren der pauschalen Messung der Genexpression sollten die Ergebnisse eines SAGE-Durchlaufes zudem repräsentativ für die betrachtete RNS-Population sein, weswegen an dieser Stelle darauf zusätzlich eingegangen werden soll.

1.6.1 Nebengütekriterien

Utilität

Die wissenschaftliche Nützlichkeit der umfassenden Genexpressionsanalyse und die Utilität von SAGE als einem Verfahren, das wesentliche Neuerungen erbringen kann und theoretisch in einem Standardlabor umsetzbar sein sollte, ist aus dem oben Gesagten evident. Im Wesentlichen besteht sie darin, daß SAGE Transkriptome ganz und digital erfassen kann, wobei bisher nicht bekannte Gene entdeckt werden können.

Ökonomie

Da SAGE im Vergleich zur Sequenzierung von ESTs wesentlich effizienter arbeitet, handelt es sich hierbei um ein relativ ökonomisches Verfahren, das nichtsdestotrotz durch seine methodische Komplexität und den hohen erforderlichen Sequenzieraufwand arbeits- und kostenintensiv ist. Dies muß jedoch in Relation zur Größe des Unterfangens, Transkriptome in ihrer Ganzheit zu analysieren, gesehen werden.

Vergleichbarkeit

Die Tatsache, daß SAGE Stichproben absolut zählt, erlaubt bei einem geregeltem Umgang mit den resultierenden Daten einen Vergleich von Expressionsprofilen, welche in verschiedenen Arbeitsgruppen erstellt wurden. Die Vergleichbarkeit von Daten wird durch das Ausmaß der Übereinstimmung beeinflußt (siehe unten). Zur Verdeutlichung: Expressionsdaten, die mittels eines relativ messenden Verfahrens wie beispielsweise dem Northern Blotting erstellt werden, bieten im Gegensatz dazu jenseits des Rahmens eines einzelnen Experimentes keine Möglichkeit des Vergleichs.

Repräsentativität

Eine andere Eigenschaft, welche von einer das Transkriptom analysierenden Methode gefordert werden muß, ist Repräsentativität. Dies impliziert, daß das beobachtete Expressionsprofil ein mit der untersuchten Boten-RNS Population sowohl qualitativ wie auch quantitativ übereinstimmendes Abbild liefert. Da SAGE jedoch mit einem zählenden und kategorisierenden Sammelverfahren verglichen werden kann und somit von wahrscheinlichkeitstheoretischer Natur ist (Man et al. 2000), kann die Repräsentativität eines SAGE Projektes lediglich in der Sprache der Wahrscheinlichkeit ausgedrückt werden. Die Repräsentativität von SAGE hängt einerseits von der in dem jeweiligen Forschungsprojekt sequenzierten Transkriptmenge ab und steigt mit deren Größe[1], andererseits von der Komplexität der Genexpression des jeweiligen Gewebes oder Zustandes. Hierbei ist zu beachten, daß in der vorliegenden Arbeit mit dem Gehirn ein Gewebe untersucht wurde, welches eine Vielzahl an unterschiedlichen Zelltypen aufweist, die wiederum viele verschiedenen Gene in Boten-RNS und Proteine umsetzen.[2] Ein repräsentatives Abbild der zugrundeliegenden Transkriptpopulation in einem derart komplexen Gewebe zu erreichen, erfordert einen höheren Sequenzierungsaufwand als für einen Zellverband, der ein einheitlicheres Zellmuster und nur wenige aktive Gene aufweist.

1.6.2 Hauptgütekriterien

Objektivität

An erster Stelle der Hauptgütekriterien steht die Objektivität eines Verfahrens, worunter verstanden wird, daß "die Registrierung der Beobachtung so weit wie möglich unabhängig von (subjektiven) Einflüssen seitens der Beobachter erfolgt" (Bortz 1990, S.39). Das heißt, daß dieses Gütekriterium lediglich den Ablese- und Bewertungsfehler involviert. Solange bei der Auswertung der sich aus der Sequenzierung ergebenden Rohdaten sowie beim Zuordnen der Genidentität allgemein anerkannte Regeln wie zum Beispiel Subtraktion der Linkersequenzen und Elimination doppelter Ditags (siehe unten) beachtet werden, kann die Objektivität von SAGE als sehr hoch erachtet werden. Bei einer derart regulierten Registrierung ist die Objektivität also evident.

---

[1] Velculescu et al. (1999) zeigen für humane Kolonkarzinomzellinien, daß die Rate der noch neu zu entdeckenden Transkripte erst ab einer sequenzierten Tagmenge von 650 000 gegen Null geht.

[2] Die Anzahl der exprimierenden Gene im Gehirn wird auf 30000 geschätzt (Michiels et al. 1999), wobei die meisten Boten-RNS Moleküle vermutlich nur selten auftreten (Sutcliffe 1988). Dies führt zu einer vergleichsweise ausgeprägten Heterogenität des zerebralen Expressionsprofils (Dokas 1983).

Reliabilität

Die Objektivität einer Beobachtung stellt eine notwendige, wenn auch nicht hinreichende Voraussetzung für das zweite Hauptgütekriterium, die Reliabilität (Zuverlässigkeit, Präzision), dar. In diese gehen zufällige Fehler einer Messung ein, welche so deren Exaktheit beeinflußen (Bortz 1990, S. 60). Da die Reliabilität von SAGE neben der methodischen Etablierung das Thema der vorliegenden Arbeit ist, wird auf die ausführliche Definition und die Implikationen dieses Gütekriteriums gesondert eingegangen werden (siehe unten).

Validität

Die Zuverlässigkeit eines Testes wiederum ist eine notwendige, wenn auch nicht hinreichende Bedingung für die Validität, die Gültigkeit, eines Resultates. Diese sagt also aus, inwiefern ein Verfahren auch das mißt, was es messen soll. Im Rahmen eines SAGE-Projektes wird die Validität von interessanten Transkripten, das heißt in der Regel von Transkripten, welche differentiell exprimiert erschienen, üblicherweise mittels eines zweiten anerkannten (klassischen) Verfahrens, welches als Außenkriterium fungiert, wie zum Beispiel der Northern Blot, eruiert.

1.7 Reliabilität

1.7.1 Reliabilität und ihre verschiedenen Komponenten

Bevor man sich mit dem Thema Reliabilität sinnvoll auseinandersetzen kann, gilt es diesen und verwandte Begriffe wie Wiederholbarkeit, Reproduzierbarkeit, Stabilität und Übereinstimmung genau zu definieren, da in der Literatur diverse Begriffe zum Thema voneinander abweichend eingesetzt werden (Daly 2000, S. 391).

Verwendung und Definition des Begriffes Reliabilität

Der Ausdruck Reliabilität wird auf zwei gänzlich unterschiedliche Weisen benutzt. Im Kontext industrieller Statistik und Qualitätskontrolle bezeichnet er die Wahrscheinlichkeit des Versagens eines Produktes als Funktion der Zeit (www.statsoft.com/textbook/stprocan.html). Das gleiche theoretische Konzept liegt der Verwendung dieses Begriffs in medizinischen Überlebensstudien zugrunde (Fisher 1993, S. 786).

In dieser Arbeit bezeichnet Reliabilität das Maß der Entsprechung von Messungen, die mit derselben Methode unter gleichbleibenden Bedingungen am selben Testmaterial durchgeführt wurden (Bortz 1990, S. 60). In diesem Sinne verwendet, mißt Reliabilität das Ausmaß, in dem

ein Testresultat repliziert werden kann. Ein Meßprozeß ist also dann als reliabel zu betrachten, wenn er bei wiederholter Anwendung konsistente Ergebnisse liefert (Daly 2000, S. 391). In dieses Konzept der Genauigkeit von Messungen gehen sämtliche zufällige[3] Fehlerquellen der Datengewinnung und -verarbeitung, der Fixation der Daten sowie der Ablesung oder Bewertung ein (Bortz 1990, S. 39). Die Ermittlung der Reliabilität einer Methode klärt nicht, ob der gemessene Wert auch der zu messende ist (Validität, Richtigkeit, Treffsicherheit - Lorenz 1996[4], S. 9), sondern bezieht sich auf die zufällige Streuung der Meßwerte (Lorenz 1996[4], S. 8).

Die verschiedenen Aspekte der Reliabilität

Es können diverse Aspekte dieses Begriffs unterschieden werden, welche das Maß der Gesamtreliabilität beeinflussen und anhand je verschiedener Untersuchungsdesigns und statistischer Verfahren bestimmt werden. Dies hat wiederum je unterschiedliche Interpretationen der Reliabilität zur Folge. Im folgenden werden diese Begriffe - Reproduzierbarkeit, Wiederholbarkeit, Stabilität und Übereinstimmung - kurz vorgestellt.

Reproduzierbarkeit ist definiert als das Ausmaß der Variabilität von Messungen, die durch den Einfluß der *Anwender* verursacht wird (www.statsoft.com/textbook/stprocan.html). In Abgrenzung zur Objektivität (subjektiver Ablese- und Bewertungsfehler) handelt es sich hierbei um zufällige Anwendungsfehler im Meßverlauf. Es kann also von einer reproduzierbaren Meßmethode gesprochen werden, wenn verschiedene Nutzer einer Meßmethode Ergebnisse liefern, die statistisch nicht signifikant differieren (www.statsoft.com/ textbook/stprocan.html).

Ob oder inwieweit eine Messung eine gute **Wiederholbarkeit** *(repeatability)* aufweist, hängt per definitionem allein von den durch die *Methode* verursachten Ungenauigkeiten ab (Altmann 1991). Überprüfbar ist dieser Aspekt der Reliabilität anhand von wiederholten Messungen durch denselben Untersucher am selben Material.

Reproduzierbarkeit und Wiederholbarkeit beziehen sich folglich grundsätzlich auf verwandte

---

[3] Das Modell, mit welchem dieser zufällige Fehler beschrieben werden kann, lautet: W = X + U, wobei die Variable W die beobachteten Werte darstellt. Die Variable X bezeichnet die wahren Werte und ist latent, das heißt, sie kann nie selbst beobachtet werden. Der gemessene Wert W weicht unter Umständen von X ab. U ist dieser als von X unabhängig betrachtete Fehler, von dem angenommen wird, daß es um einen Mittelwert von 0 streut (Spiegelman 1998). Dem gegenüberzustellen wären systematische Fehler, welche zum Beispiel durch regelhafte unterschiedliche Handhabung bestimmter Abläufe von verschiedenen Anwendern einer Methode oder durch falsche Kalibrierung von Geräten zustande kommen können. Diese beeinflussen die Validität einer Meßmethode. Gerichtete Änderungen des biologischen Materials, das heißt in unserem Fall der Genexpression, würden ebenso in einen systematischen Bias resultieren. Diesen zu messen, ist das Ziel vergleichender Untersuchungen der Genexpression.

Gesichtpunkte, allein die Ursache eventueller Variabilitäten der Meßergebnisse liegt im Fall von ersterer bei der messenden Person, während sie bei der Wiederholbarkeit der Methode selbst angelastet werden.

Weitere Begriffe, die im Zusammenhang mit Reliabilität verwendet werden, sind Verfahrens- oder Meßstabilität und Übereinstimmung.

Wenn eine Messung nach einem gewissen Zeitabstand im Sinne eines Retests am selben Testmaterial wiederholt wird, wird von der **Kurz**- und **Langzeitstabilität** (Lienert 1994) einer Messung gesprochen. Diese hängt in biologischen Zusammenhängen von der natürlichen Variabilität des "Testmaterials", der Reproduzierbarkeit und der Wiederholbarkeit einer Methode ab. Hier wird Reliabilität also aus dem Blickwinkel der zeitlichen Fluktuation definiert.

Der Begriff **Übereinstimmung** wird vor allem im Zusammenhang mit Vergleichen verschiedener Anwender eines Verfahrens oder verschiedener Methoden, die die Messung derselben Parameter zum Ziel haben, verwendet (Altman 1991). In Abgrenzung zur Reproduzierbarkeit bezieht sich dieser Benutzervergleich nicht nur auf die Genauigkeit der Anwendung einer Methode, sondern auch auf ihre Güte, das heißt ihre Validität. In dem Fall, daß es sich bei einer der beiden Methoden, welche miteinander verglichen werden, um einen etablierten Standard handelt, der als Außenkriterium verwendet werden kann, fällt die Studie in die Kategorie der Überprüfung der Validität der nicht etablierten Methode.

Reliabilität und Repräsentativität

Im Falle von SAGE ist die Frage nach der Meßgenauigkeit des Verfahrens an die Problematik seiner Repräsentativität gekoppelt, da SAGE nicht die gesamte Transkriptpopulation mißt, sondern Stichproben auszählt. Dies bedeutet, daß - selbst wenn ein SAGE Durchlauf keinerlei Meßungenauigkeiten induzieren würde - die Tagzahl für ein bestimmtes Transkript bei wiederholten Messungen im Rahmen stochastischer Wahrscheinlichkeiten schwanken würde. Eine absolute Deckung der Ergebnisse wiederholter Messungen im Rahmen einer Überprüfung der Reliabilität ist also nicht zu erwarten. Dies heißt jedoch auch, daß der Meßgenauigkeit von solch einem "sammelnden" Verfahren wie SAGE a priori bestimmte Grenzen gesetzt sind. Das bedeutet, daß die Genauigkeit in Projekten, die weniger Tags sequenzieren immer geringer ist als dies bei Projekten der Fall ist, die große Stichproben eines Transkriptoms auszählen. Erstere entnehmen eine kleinere Stichprobe und unterliegen damit größeren stochastischen Schwankung. Ganz genau zu messen (vorausgesetzt SAGE könnte das methodisch) wäre nur möglich, wenn Gesamtpopulationen ausgezählt werden würden. Da dies nicht realistisch ist, ist von SAGE keine

100%e Meßgenauigkeit zu erwarten.
Doch weshalb ist es von Interesse sich mit der Reliabilität eines Verfahrens gezielt auseinanderzusetzen? Darauf soll im nächsten Abschnitt eingegangen werden.

1.7.2 Gründe für die Messung der Reliabilität

Die Notwendigkeit einer Überprüfung der Reliabilität eines Verfahrens ergibt sich aus den folgenden drei Gründen.
Der erste ist trivial. Ein unreliabler Test wäre der Forschung wenig nützlich und eine Verschwendung finanzieller, menschlicher und tierischer Ressourcen.
Zweitens wäre im gegenwärtigen Wettstreit der verschiedenen Methoden, welche die Messung der gesamten Genexpression zum Ziel haben, derjenigen ein Pluspunkt zuzurechnen, die eine höhere Meßgenauigkeit aufweist. Um einen solchen Vergleich durchführen zu können, muß die Reliabilität jeder Methode jedoch erst einmal bekannt sein.
Drittens wird eine Einschätzung der Reliabilität benötigt, um bestimmen zu können, ob beobachtete Differenzen in der Genexpression zwischen verschiedenen Geweben oder Zuständen durch tatsächlich vorhandene Unterschiede oder von zufälligen methodischen Variabilitäten verursacht werden (Daly 2000, S. 392f).
Die Untersuchung der Reliabilität von SAGE ist also unerläßlich und deswegen Standard bei der Entwicklung eines neuen Verfahren. Im folgenden wird zusammengefaßt, inwiefern dies bereits geschehen ist.

1.7.3 Die Reliabilität von SAGE in der Literatur

Hieter und Boguski (1997) bekräftigen in ihrem Überblicksartikel zum Gebiet *'functional genomics'* die wissenschaftliche Vorgehensweise, Gütekriterien zu überprüfen, durch die Forderung nach einem reliablen Design der Verfahren, welche das Transkriptom untersuchen, um eine aussagekräftige Evaluation der Daten zu gewährleisten. Dennoch finden sich in der ersten Publikation von Velculescu et al. (1995) keine Aussagen zur Reliabilität von SAGE. In späteren Veröffentlichungen wird, wenn auf das Thema Reliabilität eingegangen wird, entweder behauptet, daß selbige vorhanden sei, ohne dies jedoch gesondert zu belegen (zum Beispiel Peters et al. 1999 oder Bertelsen und Velculescu 1998) oder es werden zur Überprüfung in einem quasiexperimentellen Setting die Häufigkeiten einiger *'Housekeeping'*- Gene herangezogen (Madden et al. 1997 und Angelastro et al. 2000a). Eine experimentelle Reliabiltätsstudie, die Expressionsprofile insgesamt vergleicht, ist nicht vorhanden.

Die Literatur zur Reliabilität von SAGE gibt also ein lückenhaftes Bild wider. Dies macht deutlich, daß in diesem Bereich noch Nachholbedarf besteht, um eine zukunftsträchtige Methode wie SAGE auf solide wissenschaftliche Fundamente zu stellen. Bevor eine solche Studie im Rahmen der vorliegenden Arbeit durchgeführt wurde, sollte eingeschätzt werden, wie sich die Reliabilität von SAGE verhalten könnte. Dazu folgt im nächsten Abschnitt die Entwicklung der Problemstellung und Hypothese.

1.8 Problemstellung und Hypothese

Gibt es eine Möglichkeit vorab einzuschätzen, ob SAGE reliabel mißt?

Diejenigen Ergebnisse, welche bisher im Rahmen von SAGE Projekten veröffentlicht wurden, erwiesen sich anhand Überprüfung einzelner Resultate durch zum Beispiel Northern Blots als valide (zum Beispiel Angelastro et al. 2000a, Chen et al. 1998, Polyak et al. 1997, Zhang et al. 1997). Ein Verfahren mit schlechter Reliabilität würde nicht gut mit einem anderen Verfahren, das als externe Kontrolle im Sinne einer Validitätsprüfung fungiert, übereinstimmen (Altman 1991, S. 401). Dies bedeutet, daß ohne ein gewisses Maß an Meßgenauigkeit keine validen Resultate im Rahmen der einzelnen Northern Blot - Kontrolle zu erwarten wären (siehe Abbildung 2). Dies leitet im Rückschluß zu der Hypothese, daß es sich bei SAGE um ein reliables Verfahren handeln müßte. Um diese Annahme zu überprüfen, ist eine explizite Untersuchung zur Abschätzung der Reliabilität notwendig und neben der praktischen Etablierung und Modifizierung der Methode sowie der Evaluierung der statistischen Auswertung von SAGE Projekten Thema dieser Arbeit.

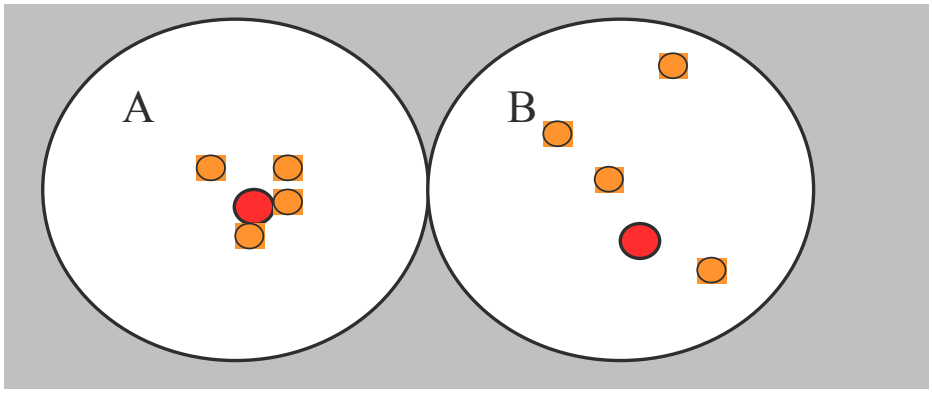

Abb. 2: Unreliable (A) und reliable (B) Messungen. In A streuen die verschiedenen Meßpunkte weit um den wahren Wert (großer Punkt), während sie in B sehr nahe beieinander und beim wahren Wert liegen (womit sie auch valide wären). Im Fall von A würde ein Vergleich mit den Ergebnissen eines Northern Blots (großer Punkt) meistens negativ ausfallen, im Fall von B positiv.

Um die verschiedenen Aspekte der Reliabilität (Reproduzierbarkeit, Wiederholbarkeit), wie sie

oben dargestellt wurden, abzuklären, wäre es notwendig, SAGE-Durchläufe in den unterschiedlichsten Konstellationen durchzuführen (Reproduzierbarkeit). Dies würde beispielsweise bedeuten, daß diverse Untersucher SAGE parallel am gleichen Material durchführen. Um genau abzuklären, welcher der Schritte von SAGE Meßungenauigkeiten induziert, wäre es notwendig, diverse Durchläufe an den verschiedenen Stufen von SAGE zu unterbrechen, das bis dahin erhaltene Material zu teilen, die Untergruppen parallel weiterzuführen und anschließend zu vergleichen.
Dies alles ist jedoch aus ökonomischen Gründen nicht möglich. Umsetzen ließ sich, mit einer Arbeitsgruppe zwei SAGE Durchläufe parallel an einer zweigeteilten Grundmenge an Boten-RNS (K1 und K2) durchzuführen und damit eine Überprüfung der Wiederholbarkeit.

Doch ist eine technisch derart komplexe Methode wie SAGE in einem molekularbiologischen universitären Standardlabor überhaupt durchführbar? Die Darstellung und Diskussion der Etablierung soll diese Frage im *ersten Teil* der Arbeit beantworten (4 Etablierung: S. 50 bis S. 111).
Im *zweiten Teil* der Arbeit (5 Statistische Evaluation und Reliabilität von SAGE) werden statistischen Aspekte von SAGE thematisiert, inklusive der Reliabilität. Da in der Literatur diverse Tests zur Berechnung der statistischen Signifikanz verwendet werden, werden diese evaluiert, um so in der Vielzahl der vorhandenen Tests eine Orientierung zu ermöglichen. Dabei werden ausgewählte Tests anhand der Daten der vorliegenden Arbeit daraufhinüberprüft, ob sie unterschiedliche Ergebnisse (Anzahl der statistisch signifikant verschiedenen Tagpaare) liefern. Dazu werden die beiden Untergruppen K1 und K2 als zu vergleichende Expressionsprofile interpretiert und die Anzahl der als statistisch verschieden ermittelten Tagpaare der untersuchten Tests miteinander verglichen. Einer der so untersuchten Test wurde im Kontext von SAGE bisher noch nicht angewandt.
Zur Ermittlung der Reliabilität werden die beiden Subgruppen K1 und K2 auf Homogenität getestet. Zusätzlich wird der Kontingenzkoeffizient zur weiteren Beurteilung der Wiederholbarkeit berechnet.
Im Rahmen der Diskussion des statistischen Abschnitts werden zuerst die für SAGE relevanten statistische Entscheidungsprozesse erläutert (S. 133ff) und die Tests aus der Literatur vorgestellt und bewertet (S. 141ff), um so Empfehlungen aussprechen zu können, welche Tests überhaupt beziehungsweise im Rahmen unterschiedlicher SAGE Projekte sinnvoll anzuwenden seien.
Im letzten Abschnitt der Diskussion (S. 155ff) werden die Ergebnisse zur Reliabilität von SAGE evaluiert und diskutiert.

2 Materialien

2.1 Kits

BigDye™ Terminator Sequencing Kit	Perkin Elmer, Vaterstetten, BRD
cDNA Synthese Kit	GibcoBRL, Eggenstein, BRD
Gene Images Kit	Amersham, Cleveland, Ohio, USA
MessageMaker™ Kit	GibcoBRL, Eggenstein, BRD
Plasmid Mini Purification Kit	Qiagen, Hilden, BRD
QIAprep 96 Turbo Miniprep Kit	Qiagen, Hilden, BRD
QIAquick Gel Extraction Kit	Qiagen, Hilden, BRD
Thermo Sequenase Sequencing Kit	Amersham, Cleveland, Ohio, USA
Trizol Reagenz	GibcoBRL, Eggenstein, BRD

2.2 Vektoren und Bakterien

pZErO™ - Vektor	Invitrogen, San Diego, CA, USA
XL1-Blue MRF Bakterien	Stratagene, La Jolla, CA, USA

2.3 Enzyme

DyNAzyme™	Biometra, Maidstone, UK
Klenow	Pharmacia, Uppsala, Sweden
Restriktionsendonukleasen und Puffer	NEB, Beverly, MA, USA
T4 Ligase und Puffer	GibcoBRL, Eggenstein, BRD
Taq-Polymerase	Perkin Elmer, Vaterstetten, BRD

2.4 Chemikalien und weitere Stoffe

Acrylamid	Serva, Heidelberg, BRD
AG 501 X-8 - Harz	BioRad, Richmond, CA, USA
Agarose	Sigma, St. Louis, MO, USA
Ammoniumacetat	Merck, Darmstadt, BRD
Ampicillin	Sigma, St. Louis, MO, USA
APS (Ammoniumpersulfat)	Serva, Heidelberg, BRD
ß-Mercaptoethanol	Sigma, St. Louis, MO, USA

Bisacrylamid	Roth, Karlsruhe, BRD
Borsäure	Roth, Karlsruhe, BRD
Bromphenolblau	Sigma, St. Louis, MO, USA
BSA (bovines Serumalbumin)	Sigma, St. Louis, MO, USA
Chloroform	Merck, Darmstadt, BRD
Calciumclorid	Sigma, St. Louis, MO, USA
DEPC (Diethylpyrocarbonat)	Sigma, St. Louis, MO, USA
Dinatriumhydrogenphosphat	Sigma, St. Louis, MO, USA
DMSO (Dimethylsulfoxid)	Sigma, St. Louis, MO, USA
DNS Größenstandards (1kb und 100bp)	GibcoBRL, Eggenstein, BRD
Dynabeads	Dynal, Oslo, Norge
Eisessig	Baker, Phillipsburg, NJ, USA
Ethanol	Merck, Darmstadt, BRD
Ethidiumbromid	Sigma, St. Louis, MO, USA
Glucose	Sigma, St. Louis, MO, USA
Glycerin	Merck, Darmstadt, BRD
Glycogen	Böhringer Mannheim, Mannheim, BRD
Glyoxallösung (40%)	BioRad, München, BRD
Isopropylalkohol	Sigma, St. Louis, MO, USA
Harnstoff	Roth, Karlsruhe, BRD
Hefeextrakt	GibcoBRL, Eggenstein, BRD
IPTG	Merck, Darmstadt, BRD
Kaliumchlorid	Merck, Darmstadt, BRD
Kaliumdihydrogenphosphat	Sigma, St. Louis, MO, USA
Lacto Tryptone	GibcoBRL, Eggenstein, BRD
LB-Flüssigmedium (Luria Broth Base)	GibcoBRL, Eggenstein, BRD
Luria Agar	GibcoBRL, Eggenstein, BRD
Magnesiumchlorid	Merck, Darmstadt, BRD
Magnesiumsulfat	Sigma, St. Louis, MO, USA
Maltose	Sigma, St. Louis, MO, USA
Mineralöl	Perkin Elmer, Vaterstetten, BRD
Na_2EDTA (Ethylendiamintetraacetat)	Serva, Heidelberg, BRD
Natriumchlorid	Roth, Karlsruhe, BRD

Natriumdihydrogenphosphat	Sigma, St. Louis, MO, USA
Natriumdodecylsulfat (SDS)	Sigma, St. Louis, MO, USA
Natronlauge	Merck, Darmstadt, BRD
Phenol	Invitrogen, San Diego, CA, USA
Phenol:Cloroform:Isoamylalkohol (25:24:1)	GibcoBRL, Eggenstein, BRD
RNS Größenstandard	GibcoBRL, Eggenstein, BRD
Salzsäure	Baker, Phillipsburg, NJ, USA
Sucarose	Sigma, St. Louis, MO, USA
SYBR Green I	Molecular Probes Inc., Eugene, OR, USA
TEMED (N, N, N‘, N‘-Tetramethylethylendiamin)	Serva, Heidelberg, BRD
Tetrazyklin	Sigma, St. Louis, MO, USA
Thiamin-HCl	Sigma, St. Louis, MO, USA
Trinatriumcitrat-Dihydrat	Merck, Darmstadt, BRD
Tris Base	Roth, Karlsruhe, BRD
Tris HCl	Roth, Karlsruhe, BRD
Tween 20	Sigma, St. Louis, MO, USA
Zeozin	Invitrogen, San Diego, CA, USA

2.5 Puffer, Lösungen und Medien

Die verwendeten Puffer, Lösungen und Medien und ihre Herstellung sind in Tabelle 3 dokumentiert.

Lösung	Herstellung
2 x B&W Puffer	10 mM TrisHCl (pH 7,5), 1 mM EDTA und 2 M NaCl in Aqua bidest. lösen und bei Raumtemperatur lagern.
DEPC - Wasser	0,1% Diethylpyrocarbonat in Aqua bidest für 24 Stunden inkubieren und anschließend autoklavieren.
DNS-Probenpuffer	0,25% Bromphenolblau und 40% Sucrose in Wasser, Lagerung bei 4°C.
Ethidiumbromid	10 mg/ml Stocklösung, Lagerung bei 4°C.
Glyoxal	40%ige Glyoxallösung (6M) mit AG 501 X-8 (Biorad, Richmond, CA, USA) deionisieren bis der pH größer als 5 ist Lagerung der Aliquots bei -20°C.
LB-Agar-Platten mit geringer Salzkonzentration und Zeocin™	10g Trypton, 5g Hefeextrakt und 5g NaCl in 950ml Aqua bidest. lösen. Den pH-Wert mit 5M NaOH auf 7,5 einstellen, 15g Agar zugeben und das Volumen auf 1 l ergänzen. Nach Autoklavierung den Agar auf 50°C abkühlen lassen und 500µl Zeocin™ (Stockkonzentration: 100mg/ml) sowie IPTG (Endkonzentration: 1 mM) zugeben.
LB (Luria - Bertani) - Medium	1 x LB-Lösung vor Gebrauch ansetzen (25 g auf 1 l H_2O) und sofort autoklavieren.

LoTE	3 mM TrisHCl (pH 7,5) und 0,2 mM EDTA (pH 7,5) in Aqua bidest. lösen und bei 4°C lagern.
Luria Agar - Platten	37 g Agar in 1 l H_2O lösen und sofort autoklavieren. Sobald der Agar auf circa 50°C abgekühlt ist, unter der Sterilbank das Antibiotikum hinzugeben und die Platten gießen. Den Agar aushärten lassen und die Platten umgekehrt bei 4°C lagern.
M9 - Medium	6g Na_2HPO_4, 3g KH_2PO_4, 1g NH_4Cl und 0,5g NaCl mit Wasser auf 1l ergänzen, autoklavieren und auf 50°C abkühlen lassen. 200ml dieses konzentrierten Ansatzes mit sterilem deionisiertem Wasser auf 1l ergänzen. Zugabe von 10ml der Filter-strilisierten Mischung von 2ml $MgSO_4$ (1M), 2g Glucose, 0,1ml $CaCl_2$ (1M) und 1ml Thiamin-HCl (1M) in Wasser.
M9 Agar - Platten	Vor Zugabe der der Filter-sterilisierten Lösung zu den M9 Salzen zu diesen 15g Agar zugeben und autoklavieren. Später den M9-Agar aushärten lassen und die Platten umgekehrt bei 4°C lagern.
0,1 M Natriumphosphatpuffer (Northern - Blot)	4,23 ml 1M NaH_2PO_4 und 5,77 ml 1M Na_2HPO_4 mit Aqua bidest. auf 100 ml auffüllen, den pH kontrollieren und autoklavieren.
PC8	480 ml Phenol (65°C warm), 320 ml 0,5M Tris-HCl (pH 8) und 640 ml Chloroform ansetzen und schütteln. Für 2 bis 3 Stunden bei 4°C kühlen. Erneut schütteln und nochmals für 2 bis 3 Stunden bei 4°C kühlen. Wässrige Phase entfernen. Aliquotieren und bei -20°C lagern.
10 x PCR Puffer	166 mM $(NH_4)_2SO_4$, 670 mM Tris (pH 8,8), 67 mM $MgCl_2$ und 100 mM B-Mercaptoethanol ansetzen. In 0,5 ml Portionen aliquotieren und bei -20°C lagern.
20 x SSC	175,3 g NaCl (entspricht 3 M) und 88,2 g $Na_3Citrat{\cdot}H_2O$ (entspricht 0,3 M) in 800 ml Aqua bidest lösen, den pH auf 7,0 einstellen und auf 1000 ml auffüllen.
SOB - Medium	20 g Trypton, 5 g Hefeextrakt und 0,5 g NaCl in 950 ml Aqua bidest. lösen. 10 ml einer 250 mM KCL Stock-Lösung (1,86 g KCL in 100 ml Wasser) zugeben. pH Einstellung mit 5 M NaOH auf 7,5 und Ergänzung des Volumens auf 1 l. Autoklavieren, auf circa 55 °C abkühlen lassen und 10 ml sterile 1 M $MgCl_2$ zugeben. Falls erforderlich, ein Antibiotikum (zum Beispiel Zeocin einer Endkonzentration von 50 µg/ml oder Tetrazyklin einer Endkonzentration von 12,5µg/ml) zugeben. Bei 4°C lagern.
SOC - Medium	Zum 1l SOB Medium (ohne Antibiotikum) 7,2ml 50% Gluose hinzugeben.
50 x TAE Puffer	242g Tris Base, 57,1ml Eisessig und 100ml 0,5 M EDTA (pH 8) mit Aqua bidest auf 1l ergänzen und autoklavieren.
5 x TBE Puffer	54g Tris base, 27,5g Borsäure und 20ml 0,5M EDTA (pH 8) mit Aqua bidest auf 1l ergänzen und autoklavieren.
TE - Puffer (pH 8)	10mM Tris (pH 8) und 1mM EDTA (pH 8) und autoklavieren.
1 M Tris HCl	121,g Tris Base auf 1 l H_2O vor dem Auffüllen mit konzentrierter HCl den pH einstellen: für einen pH von 7,4 mit 70ml HCl, für einen pH von 7,6 mit 60ml HCl und für einen pH von 8 mit 42ml HCl.
YT - Medium mit Zeocin™	Um 1l dieses Mediums herzustellen, wurden 8g Trypton, 5g Hefeextrakt und 2,5g NaCl in 900ml Aqua bidest. gelöst. Nach Einstellung des pHs auf 7,5 mit 5M NaOH wurden 15g Agar zugegeben und das Volumen auf 1l ergänzt. Nach der Autoklavierung den Agar auf 50°C abkühlen lassen und 500µl Zeocin™ (Stockkonzentration: 100mg/ml) zugeben.

Tabelle 3. Puffer und Lösungen.

2.6 Nukleotide, Linker und Primer

Die verwendeten (Oligo)Nukleotide sind in Tabelle 4 dokumentiert.

Nukleotid	Sequenz (5' - 3')	Position	Literatur	Firma
dNTP (beziehungs-weise dATP, dCTP, dGTP, dTTP)	Einzelnukleotide	-	-	Amersham, Cleveland, Ohio, USA
7-deaza-dGTP	-	-	-	Boehringer Mannheim, Mannheim, BRD
Fluoreszein-12-dUTP	-	-	-	DuPont NEN, Bad Homburg, BRD
Linker 1 A	TTTGGATTTGCTGGTGCAGTACCACTAGG CTTAATAGGGACATG	keine	Velculescu et al. SAGE - Detailed Protocol 1.0c	Pharmacia, Uppsala, Sweden
Linker 1 B	TCCCTATTAAGCCTAGTTGTACTGCACCA GCAAATCC [mit Aminomodifikation des C7]	keine	Velculescu et al. SAGE - Detailed Protocol 1.0c	Pharmacia, Uppsala, Sweden
Linker 2 A	TTTCTGCTCGAATTCAAGCTTCTAACGAT GTACGGGGACATG	keine	Velculescu et al. SAGE - Detailed Protocol 1.0c	Pharmacia, Uppsala, Sweden
Linker 2 B	TCCCCGTACATCGTTAGAAGCTTGAATTC GAGCAG [mit Aminomodifikation des C7]	keine	Velculescu et al. SAGE - Detailed Protocol 1.0c	Pharmacia, Uppsala, Sweden
Primer 1	GGATTTGCTGGTGCAGTACA	keine	Velculescu et al. SAGE - Detailed Protocol 1.0c	Eurogentec, Seraing, Belgium
Primer 2	CTGCTCGAATTCAAGCTTCT	keine	Velculescu et al. SAGE - Detailed Protocol 1.0c	Eurogentec, Seraing, Belgium
M13 Forward Primer (F)	GTAAAACGACGGCCAGT	M13	Velculescu et al. SAGE - Detailed Protocol 1.0c	ABI Perkin Elmer, Vaterstetten, BRD
M13 Reverse Primer (R)	GGAAACAGCTATGACCATG	M13	Velculescu et al. SAGE - Detailed Protocol 1.0c	ABI Perkin Elmer, Vaterstetten, BRD

AKAP B5IAT7 Rev	TTGTGTTGTAGCTGTCTACCCACTG	Akap149 (2828 - 2804)	EMBL Database Acc. No. X97335	MWG Biotech, Ebersberg, BRD
AKAP IIB5EF	CTGAACCTCACCAATATCTACGCTC	Akap149 (2156 - 2180)	EMBL Database Acc.No. X97335 (Trendelen-burg et al. (1996)	MWG Biotech, Ebersberg, BRD

Tabelle 4. Nukleotide, Linker und Primer.

2.7 Geräte und Hilfsmaterial

- 3MM-Papier — Whatman, Maidstone, UK
- ABI 373A Sequenzierautomat — Perkin Elmer, Vaterstetten, BRD
- *ALFexpress* Sequenzierautomat — Pharmacia, Uppsala, Sweden
- Autoklav Varioklav — H + P Labortechnik, Oberschleißheim, BRD
- Brutschrank B5060 EC-CO2 — Haereus, München, BRD
- Elektroporationsapparat *E.coli Pulser* — BioRad, Richmond, CA, USA
- Fotopapier Color Video Print Typ S — Hitachi, Tokyo, Japan
- Gelelektrophoreseapparatur horizonal — BioRad, Richmond, CA, USA
- Gelelektrophoreseapparatur vertikal PSD9S — PEQLABS, Erlangen, BRD
- Gewebshomogenisator — Falcon, Heidelberg, BRD
- Inkubator 60°C-200°C U 40 — Memmert, Schwabach, BRD
- Inkubator Plus II — Gallenkamp, Loughborough, UK
- Küvetten für *E.coli Pulser* — BioRad, Richmond, CA, USA
- Medienflaschen — Falcon, Heidelberg, BRD
- Mikrowelle — Bosch, BRD
- Nylonmembran — Böhringer Mannheim, Mannheim, BRD
- PCR Theromcycler (Trio-Thermoblock) — Biometra, Maidstone, UK
- Petrischalen — Falcon, Heidelberg, BRD
- Pipetten — Labsystems, Helsiniki, Finland
- Photoapparatur / Kamera CU-5/75mm — Polaroid, Hertfordshire, UK
- Reaktionsgefäße 1.5, 2, 15ml — Eppendorf, Hamburg, BRD
- Reaktionsgefäßschüttler 5432 — Eppendorf, Hamburg, BRD
- Röntgenfilme — Kodak, Stuttgart, BRD
- Schüttler REAX 2000 — Heidolph, Schwabach, BRD

• Schüttelwasserbad Polytest 20	Bioblock Scientific, Illkirch Cedex, France
• Spektralphotometer PM 4	Zeiss, Jena, BRD
• Sterilbank Bio48	Faster, Berlin, BRD
• SYBR Green Filter	Molecular Probes Inc., Eugene, OR, USA
• Tischzentrifuge Typ 1 - 13	Sigma, St. Louis, MO, USA
• Ultrazentrifuge *Centricon* T2050	Kontron, Zürich, CH
• UV-Bildschirm	Bioblock Scientific, Illkirch Cedex, France
• Waage	Eppendorf, Hamburg, BRD
• Zentrifugationssäulen (*SpinX*)	Costar Inc., Cambridge, MA, USA
• Zentrifuge 5804 R	Eppendorf, Hamburg, BRD

2.8 Computer und Software

Zur Erstellung der vorliegenden Arbeit wurde ein PC (Network) mit Intel Pentium II Prozessor (300MHz) verwendet.

Folgende Programme wurden eingesetzt:

MS Office 1997 / 2000 Professional Edition	Microsoft, Redmond, USA
SPSS Version 10.0	SPSS GmbH, München, BRD
S-Plus Version 6.0	Insightful, Seattle, WA, USA
SAGE 300 Version 3.01	Baltimore, MD, USA (Zhang et al. 1997)
SAGEstat	Amsterdam, Niederlande (Kal et al. 1999)
Primer 1.01	Scientific & Educational Software, Durham, NC, USA
Mathematica 4.1	Wolfram Research Europe, Oxfordshire, UK

3 Methoden

3.1 RNase freies Arbeiten

Um die Degradierung durch ubiquitär vorhandene RNasen beim Arbeiten mit RNS zu umgehen, wurden bei 180°C für mindestens acht Stunden sterilisierte Glasgefäße sowie fabrikneue Reaktionsgefäße, Pipetten und Chemikalien benutzt. Sämtliche Reagenzien wurden RNase-frei angesetzt. Lösungen, welche keine Substanzen mit reaktiven Aminogruppen enthielten, wurden über Nacht mit 0,1% DEPC inkubiert und anschließend autoklaviert. Elektrophoreseapparaturen wurden gründlich mit handelsüblichen Geschirrspülmitteln gereinigt, mit 10% SDS behandelt und mit RNase freiem Wasser gespült. Außerdem wurden die Arbeiten mit Handschuhen durchgeführt.

3.2 Steriles Arbeiten

Um beim Arbeiten mit Reinkulturen Kontaminationen mit anderen Organismen zu verhindern, wurden Nährlösungen und Gerätschaften im Autoklaven bei 1bar Überdruck und 121°C für zwanzig Minuten sterilisiert beziehungsweise steril erhältliche Materialien verwendet. Arbeiten, welche Sterilität erforderten, wurden unter einem gesonderten Abzug durchgeführt.

3.3 Phenol – Choroform – Extraktion

Zur Entfernung von Proteinen und hydrophoben Stoffen aus wäßrigen DNS-Lösungen wurden diese auf Eis mit dem gleichen Volumen eines Phenol-Chloroform-Isoamylalkohol-Gemisches (25:24:1) versetzt und vermischt. Die Phasentrennung erfolgte durch zehnminütige Zentrifugation (Eppendorf Zentrifuge 5804 R) bei 4°C mit 15000 x g. Die obere wäßrige Phase, welche die DNS enthält, wurde abgenommen und mittels Ethanol gefällt, während die Interphase und die phenolische Phase verworfen wurden.
Alternativ wurde ein Gemisch aus Phenol, 0,5M TrisHCl - Puffer (pH 8) und Chloroform im Verhältnis 3:2:4 (PC8) verwendet. Bei ansonsten gleicher Vorgehensweise wurde hier lediglich zwei Minuten lang bei 4°C mit 6.000 x g zentrifugiert.

3.4 3. 4 Ethanolpräzipitation von Nukleinsäuren

Zur Fällung von Nukleinsäuren in wäßrigen Lösungen wurden soweit nicht anders angegeben die Ansätze auf 0,3M Ammoniumacetat eingestellt, mit 60µg Glycogen und zweifachem Volumen kalten 100% Ethanol versetzt und für mindestens eine Stunde bei -20°C inkubiert. Die gefällte DNS beziehungsweise RNS wurde für zehn Minuten bei 6.000 x g (4°C) zentrifugiert. Der Überstand wurde abgegossen, das Sediment zweimal mit kaltem 75% Ethanol gewaschen,

einige Minuten bei Raumtemperatur getrocknet und in LoTE - Puffer aufgenommen.
Bei der in einigen Schritten von SAGE verwendeten sogenannten hoch konzentrierten Fällung wurden die Nukleinsäure-Lösungen auf 0,8M Ammoniumacetat eingestellt, ebenfalls mit 60µg Glycogen und hier mit dem dreifachem Volumen kaltem Ethanol versetzt. Anschließend wurde wie oben vorgegangen.

3.5 Bestimmung von Nukleinsäurenkonzentrationen

3.5.1 Messung der optischen Dichte

Im Photometer erfolgte die Messung der optischen Dichte (OD) bei 280nm (OD_{280}) und bei 260nm (OD_{260}). Der Grad der Verunreinigung mit Proteinen wurde aus dem Quotienten OD_{280}/OD_{260} bestimmt, wobei ein Wert von größer 1,5 angestrebt wurde. Bei der üblicherweise für die Messungen verwendeten Verdünnung von 1:100 ergab sich die Konzentration K der Nukleinsäuren nach folgender Formel: $K\ [\mu g/ml] = OD_{260} \times 100 \times c$. Für c wurde bei Messungen von Doppelstrang-DNS 50µg/ml, 37µg/ml bei Einzelstrang-DNS, im Falle von RNS 40µg/ml (Einzelstrang-RNS) eingesetzt.

3.5.2 Semiquantitative DNS Mengenbestimmung mit Ethidiumbromid

Aus DNS-Marker (Konzentration der Stocklösung: 500ng/µl) wurden folgende DNS-Standards hergestellt: 0ng/µl, 1ng/µl, 2,5ng/µl, 5ng/µl, 7,5ng/ml, 10ng/µl und 20ng/µl. Von der zu quantifizierenden DNS-Lösung wurde aus 1µl 1:5, 1:25, 1:125 und 1:625 Verdünnungen hergestellt. Zu jeweils 4µl der Standards und der Proben wurde das gleiche Volumen an 1µg/ml Ethidiumbromid zugegeben. Auf einer auf einem UV-Bildschirm plazierten Plastikfolie wurden sämtliche Lösungen punktförmig aufpipettiert und fotographiert. Durch den visuellen Vergleich der Intensitäten von Standards und Proben wurde die DNS Konzentration der Proben geschätzt.

3.6 Gelelektrophorese

3.6.1 Agarosegelelektrophorese

Die Auftrennung von DNS beziehungsweise RNS nach ihrer Länge erfolgte im Agarosegel in einer horizontalen Gelelektrophoreseapparatur, wobei die Agarosekonzentration entsprechend dem gewünschten Trennbereich zwischen 0,7% und 3% lag. Nach Lösen der Agarose (Sigma, USA) in 1x TAE-Puffer oder 1x TBE-Puffer durch Aufkochen in einem Mikrowellengerät wurde sie auf mindestens 60°C abgekühlt und mit Ethidiumbromid (Endkonzentration 1µg/ml) versetzt. Nach Gießen und Verfestigen des Gels wurden die mit entsprechendem Puffer versetzten Nukleinsäurengemische aufgetragen. Für den Lauf wurde eine Gleichspannung von

ein bis fünf Volt/cm über einen Zeitraum von ein bis zwei Stunden angelegt. Anschließend wurde das Gel auf einem UV - Schirm fotographiert.

3.6.2 Polyacrcylamidgelelektrophorese

Im Rahmen bestimmter SAGE - Schritte wurde je nach erwünschtem Trennbereich ein 18%, 12%, 8% oder 6% Polyacrylamidgel verwendet. Diese Elektrophorese fand in einer vertikalen Elektrophoreseapparatur (PEQLAB PSD9S - Kammer mit 1,5 mm Abstandhaltern) statt. Als Laufpuffer wurde 1x TAE-Puffer verwendet.

Für ein 12% Gel wurde folgende Zusammensetzung verwendet (Adjustierung der Polyacrylamidanteile je nach gewünschter Konzentration):

- 10,5ml 40% Polyacrylamid (19:1 Acrylamid : Bisacrylamid)
- 23,5ml destilliertes H_2O
- 700µl 50 x Trisazetatpuffer
- 350µl 10% APS
- 30µl TEMED

Nach mindestens einstündigem Polymerisieren des Gels wurde der Kamm vorsichtig entfernt und die ebenfalls mit entsprechendem Puffer versehenen Proben aufgetragen. Anschließend wurde die Elektrophorese bei 90 bis 160 V drei Stunden durchgeführt, wobei mit niedriger Spannung (90 V) begonnen wurde, diese nach einer halben Stunde auf 110 V und nach insgesamt einer Stunde auf 130 V gesteigert wurde. Während des gesamten Laufs wurde auf 10°C gekühlt.

Um die DNS sichtbar zu machen, wurde das Gel nach Beendigung der Elektrophorese, das heißt sobald das Bromphenolband das Ende des Gels erreicht hatte, für zwanzig Minuten in einer Färbelösung, welche aus 65ml 1x TAE-Puffer und 6ml SYBR Green I (Molecular Probes Inc.) einer 1:10000 Verdünnung bestand, bei Raumtemperatur inkubiert. Zur Dokumentation wurde das Gel auf einem UV - Schirm unter Verwendung eines SYBR Green Filters (Molecular Probes Inc.) fotographiert.

3.7 Restriktionsenzymverdau

Zum enzymatischen Verdau von DNS wurden die DNS-Lösungen (1 - 20 µg DNS) mit biszu 100µg/ml BSA, dem zum Enzym passenden Restriktionspuffer und mit der entsprechenden Menge des Enzyms versetzt. Bei dem Verdau mit BsmFI waren dies 4U, mit NlaIII 17 (1. Verdau) oder 25 (2. Verdau) U/µg DNS. Nach Inkubation für eine Stunde bei 37°C beziehungsweise 65°C wurde eine Phenol - Chloroform - Extraktion und eine Ethanolfällung durchgeführt. Wenn eine Analyse des Verdaus von Interesse war, wurde diese per

Gelelektrophorese ermöglicht.

3.8 Begradigen von cDNS

Um Überhänge von verdauter doppelsträngiger cDNS zu beseitigen und stumpfe Enden herzustellen, wurde in einem 50µl Ansatz 10µl der gereinigten DNS-Lösung (circa 2 - 3µg) mit der entsprechenden Menge (5µl) des Zweitstrangsynthese Puffers (BRL cDNA Synthese Kit), 1µl 100 x BSA, 1µl einer dNTP Mischung (25mM; Amersham) und 3U Klenow (Pharmacia) versetzt. Nach einer Inkubation von dreißig Minuten bei 37°C wurde eine Phenol - Chloroform - Extraktion und eine Ethanolfällung durchgeführt.

3.9 Ligation

Zur Ligation zweier DNS - Sequenzen wurden im 10µl Ansatz die DNS-Lösungen (mehrere hundert Nanogramm bis 4µg) mit 2 µl 5 x T4 Ligase Puffers und T4 Ligase (GibcoBRL) (Endkonzentration: 1U/µl) versetzt. Die Inkubation bei 16°C erfolgte über Nacht, wenn DNS Stücke in einen Vektor geklont wurden, oder zwischen zehn Minuten und zwei Stunden lang im Falle der Ligation zweier DNS-Stückchen.

Im Fall der Ligation der SAGE-Linker an cDNS-Fragmente wurde jeweils die gesamte Menge der immobilisierten cDNS und 2µg von einem der beiden Linkerduplexe (A beziehungsweise B) eingesetzt. Nach Inkubation bei 50°C für zwei Minuten und anschließend bei Raumtemperatur für eine Viertelstunde wurde jedem Ansatz 10 Units T4 Ligase zugegeben und bei 16°C für zwei Stunden inkubiert. Nach der Ligation wurde viermal mit 200µl B&W Puffer und zweimal mit 200µl 1x NEB Puffer gewaschen.

3.10 Präparation kompetenter Bakterien und Anlegen einer Bakterienstammsuspension

Auf M9-Platten wurde ein Einzelaustrich von XL1-Blue MRF Bakterien (Stratagene) (mit 12,5µg/ml Tetrazyklin) angelegt. Nach Inkubation dieses Ausstrichs bei 37°C über Nacht wurde als Vorkultur 5 ml LB-Medium (mit 0,2% Maltose, 10mM $MgSO_4$, 12,5µl/ml Tetrazyclin) mit einer einzelnen Kolonie inokuliert und erneut über Nacht bei 37°C inkubiert. Für die Hauptkultur wurde damit 250 ml 2 x YT- Mediums (inklusive 1 ml Tetrazyclin der Konzentration 5 mg/ml) inokuliert und für vier Stunden bei 37°C inkubiert. Nach zehnminütigem Kühlen auf Eis wurde die Bakterienkultur bei 4000 x g (4°C) für zehn Minuten zentrifugiert. Der Überstand wurde verworfen und das Bakterienpellet in kaltem sterilem Wasser gelöst. Dieses Vorgehen wurde mit absteigenden Volumina Wasser sechsmal wiederholt, wobei ab dem drittenmal die Bakterien in 10%igem kalten anstelle von Wasser gelöst wurden. Die gereinigten Bakterien wurden in 90 µl

10% sterilem Glyzerol aufgenommen und entweder sofort weiterverwendet oder mittels flüssigem Stickstoff eingefroren und bei -80°C aufbewahrt.

3.11 Transformation per Elektroporation

Zur Elektroporation kompetenter Bakterien wurde 40µl Bakterienlösung mit 1µl der zu sequenzierenden DNS-Lösung versetzt. Nach Inkubation auf Eis für eine Minute wurde die Mixtur in einem Elektroporationsapparat (Bio Rad E.coli Pulser) einem Strompuls (12,5 kV/cm, 200 Ohm, 25µF Kapazität, 4ms) ausgesetzt und anschließend sofort mit 1ml SOC-Medium versetzt. Nach einstündiger Inkubation bei 37°C wurden die Bakterien auf Low Salt LB-Agar-Platten, welche IPTG und Antibiotika (Zeocin und Tetrazyclin) enthielten, ausplattiert und über Nacht bei 37°C wachsen gelassen. Bei Verwendung des Vektors pZErO™ - 1 (Invitrogen) konnten so nur Kolonien entstehen, deren Bakterien den die entsprechende Antibiotikaresistenz tragenden Vektor aufgenommen hatten, wobei wiederum lediglich Bakterien überlebensfähig waren, welche Vektoren mit Fremd-DNS aufwiesen, da ohne DNS-Insert ein letal wirkendes Protein (CcdB) gebildet wurde (doppelte positive Selektion). Die Zugabe von Tetrazyklin sollte das Wachstum anderer Bakterien als der zur Transfektion verwendeten verhindern. Die entstandenen Kolonien konnten einer Kontroll-PCR unterzogen werden und wurden dann einzeln gepickt und in 2ml SOB-Medium, das Zeozin enthielt, über Nacht bei 37°C vermehrt. Es schloß sich eine Reinigung der Plasmide an.

3.12 Gesamt-RNS Isolierung aus Gewebe

Die Isolierung von RNS aus Gewebe erfolgte mit *Trizol*® entsprechend dem von GibcoBRL mitgelieferten Protokoll. Pro 50mg Gewebe wurde 1ml *Trizol*® zur Homogenisation verwendet, was 8 ml Reagenz pro Maushirn entsprach. Nach fünfminütiger Inkubation des homogenisierten Materials bei Raumtemperatur wurde pro ml verwendeten *Trizol*s® 200µl Chloroform dazugegeben und erneut für drei Minuten bei Raumtemperatur inkubiert. Nach Zentrifugation bei 12.000 x g (4°C) für 15 Minuten wurde die obere wäßrige Phase, in welcher die RNS gelöst war, abgenommen, während die phenolische und die Interphase verworfen wurden. Die Fällung der RNS erfolgte mittels Zugabe von 500µl Isopropylalkohols und 0,5 - 1µl RNase freien Glycogens pro anfangs verwendeten ml *Trizol*® durch zehnminütige Inkubation bei Raumtemperatur und anschließender Zentrifugation bei 12.000 x g (4°C) für zehn Minuten. Nachdem das RNS-Pellet anschließend einmal mit 75% kaltem Ethanol gewaschen und bei Raumtemperatur getrocknet worden war, wurde es in LoTE aufgenommen. Um nachzuweisen, daß die RNS in Verlauf der Isolierung nicht degradiert worden war, wurde in einem 1% Agarosegel eine Elektrophorese durchgeführt. Die Konzentration der RNS-Lösung wurde per

Messung der optischen Dichte bestimmt.

3.13 Boten-RNS - Präparation mittels *Message Maker*™ Kit

Für die Synthese von Boten-RNS wurde der *Message Maker*™ Kit (Gibco/BRL) nach Angaben des Herstellers verwendet. Um die Entfernung der ribosomalen RNS zuverlässig zu gewährleisten, wurde die Konzentration der isolierten Gesamt-RNS auf 0,55mg/ml oder geringer eingestellt. Nach Denaturierung der RNS bei 60°C für zwanzig Minuten und folgender Inkubation auf Eis für fünf Minuten wurde die NaCl - Konzentration der Proben auf 0,5M adjustiert. Zur Bindung der Boten-RNS wurde diese mit einem der Menge an Gesamt-RNS entsprechendem Volumen einer Oligo(dT) Cellulose - Suspension (zum Beispiel 1ml bei 0,5 - 1mg an Gesamt-RNS) bei 37°C unter mehrmaligem Invertieren für 30 Minuten inkubiert. Im Anschluß daran wurde die RNS - Oligo(dT) Cellulose Lösung mit den mitgelieferten Waschpuffer in einer Filterspritze zweimal gewaschen. Zum Lösen der Boten-RNS wurde 65°C warmes destilliertes Wasser verwendet, die folgende Zentrifugation bei 2.600 x g (4°C) für drei Minuten diente der Entfernung restlicher Cellulosepartikel. Der die Boten-RNS enthaltende Überstand wurde einer Ethanolfällung über Nacht unterzogen, wobei in diesem Fall 50µg/ml RNS-Lösung an Glycogen, ein Zehntel des Gesamtvolumens an 7,5M Ammoniumacetat und das zweifache Volumen an -20°C kaltem Ethanol verwendet wurden. Nach Zentrifugation bei 2.600 x g (4°C) für dreißig Minuten wurde die Boten-RNS in LoTE aufgenommen und ihre Konzentration per Messung der optischen Dichte bestimmt. Die Qualität der Boten-RNS wurde per Agarosegelelektrophorese und Northern Blot auf DNS Kontamination und Degration der RNS überprüft.

3.14 cDNS - Herstellung

Die Herstellung von biotinylierter cDNS erfolgte entsprechend dem von GibcoBRL mit dem *cDNA Synthesis System* mitgelieferten Protokoll. Die Herstellung des ersten DNS-Stranges aus mindestens 5µg Boten-RNS erfolgte in einen 50µl Ansatz mit folgenden Komponenten auf Eis :

- 10µl 5 x Erststrangpuffer, was eine Endkonzentration von 50mM Tris-HCl (pH 8,3), 75mM KCl und 3mM $MgCl_2$ ergab.
- 2,5µl einer 10mM Nukleotid-Mischung
- 4µl des PAGE gereinigten Oligo$(dT)_{20}$-5' Biotin-Primers (2,5µg) aus einer Stocklösung der Konzentration 100pmol/µl
- 10µl Boten-RNS (5µg)
- 5µl 0,1M DTT

Das Volumen wurde mit DEPC behandeltem Wasser auf 47,5µl ergänzt. Nach Inkubation bei

65°C für zwei Minuten und anschließend für den gleichen Zeitraum bei Raumtemperatur wurde der Ansatz mit 2,5µl M-MLV Reverse Transkriptase, was einer Endkonzentration von 10000 Einheiten/ml entsprach, versetzt, erneut inkubiert (eine Stunde bei 37°C) und anschließend auf Eis gestellt. Falls lediglich eine einsträngige und nicht-biotinylierte cDNS erwünscht war (zum Beispiel als Matrize für die Herstellung von Sonden), wurde zu diesem Zeitpunkt die Reaktion durch Zugabe von 1µl 0,25M Na_2EDTA (pH 7,5) beendet. Außerdem wurde hierzu anstelle des biotinylierten Primers ein einfacher Oligo (dT)- Primer verwendet.

Zur Synthese des zweiten Stranges wurden zu obiger Probe auf Eis folgende Reagenzien hinzugegeben:

- 40µl des 10 x Zweitstrangpuffer, was zu einer Endkonzentration von 25mM TrisHCl (pH 8,3), 100mM KCl, 10mM $(NH_4)_2SO_4$, 5mM $MgCl_2$, 0,15mM NAD und 5mM DTT führte.
- 7,5µl der 10mM Nukleotid-Mischung
- 10µl *E. coli* DNA Polymerase I (250U/ml)
- 1,75µl *E. coli* RNase H (8,5U/ml)
- 1,25µl *E. coli* DNA Ligase (30U/ml)
- 289,5µl DEPC-behandeltes Wasser

Nach Inkubation bei 16°C für zwei Stunden wurde die Probe mit 25µl 0,25M Na_2EDTA (pH 7,5) versetzt, die cDNS mittels Phenol/Chloroform extrahiert, mit Ethanol gefällt und in 21µl LoTE aufgenommen. Zur Überprüfung der Qualität der cDNS wurde mit 1µl der Lösung eine Agarosegelelektrophorese durchgeführt. Des weiteren wurde die Konzentration der cDNS mittels Ethidiumbromid Punktquantifikation (siehe S. 34) abgeschätzt sowie ein cDNS Southern - Blot durchgeführt.

3.15 Polymerasekettenreaktion (PCR)

3.15.1 Allgemeine Prinzipien

Die Optimierung der PCR erfolgte standardmäßig in einem 50µl Ansatz aus folgenden Komponenten:

- 5µl einer 10 x Probenpuffers, was für die meisten Polymerasen eine Mg^{2+}-Konzentration von 1,5mM im Ansatz ergab.
- Nukleotid-Gemisch, welches zu einer Endkonzentration eines jeden dNTPs von 200µM führte.
- je 25pmol der zwei Primer, was einer Endkonzentration von je 500µM entspricht.
- 5 Units der Polymerase (Taq-Polymerase oder DyNAzyme™)
- circa 10ng der Matrizen-DNS
- das entsprechende Volumen Wasser, um das Gesamtvolumen 50µl zu erreichen.

Zur Amplifikation von Plasmid-DNS und spezifischer SAGE-DNS-Stücke wurde dem Reaktionsansatz DMSO zugefügt.
Zur Amplifikation der SAGE-Ditags wurden für einen 50µl Standardansatz folgende Reagenzien verwendet:

- 31,5µl Aqua
- 3µl DMSO
- 5µl 10 x PCR Puffer
- je 0,5µl der Primer A und B (100pmol/µl)
- 7,5µl dNTPs (10mM)
- 1µl Taq (5U/µl)(Perkin Elmer)
- 1µl der Ditag-Verdünnungen

Für die Screening-PCRs der klonierten Tagketten mit M13 F (universal) und R (reverse) Primern wurde folgender 25µl Ansatz gewählt:

- 2,5µl10 x PCR Puffer
- 1,5µl DMSO (Sigma)
- 0,5µl dNTPs (10mM)
- 2µl M13 F - Primer (5pmol/µl)
- 2µl M13 R - Primer (5pmol/µl)
- 16,25µl Aqua
- 0,25µl Taq (5U/l)(Perkin Elmer)

Sämtliche Proben wurden zum Schutz gegen Volumenschwankungen durch Verdunstung mit circa 30µl Öl überschichtet, zusätzlich wurde stets die Deckelheizung des Thermocyclers (Trio-Thermoblock, Biometra) eingeschaltet.
Um eine größere Spezifität und Sensitivität zu erreichen, wurde eine Warmstart-PCR durchgeführt, das heißt, daß die Polymerase erst nach einem ersten Denaturierungsschritt bei 95°C von zwei Minuten auf Eis dem Ansatz zugegeben wurde. Aus diese Weise wurde beispielsweise eine unspezifische Anlagerung von Primern bei niedrigeren Temperaturen und die anschließende Verlängerung - bei bereits vorhandener Polymerase - vermindert. Zur Gewinnung von Amplifikaten mit möglichst wenigen Sequenzfehlern wurde bei der Herstellung von Sonden (Northern und cDNS Southern Blot) die Polymerase DyNAzyme™ (Biometra) verwendet, welche nach Herstellerangaben eine wesentlich geringere Fehlerrate als die Taq DNA

Polymerase aufweist.

Bei unbefriedigenden Resultaten der Standard-PCR wurde diese durch Variation der einzelnen Parameter wie Konzentration, Temperatur, Länge der einzelnen Schritte im Verlauf der Amplifikation und Anzahl der Zyklen optimiert. Erfahrungsgemäß spielt die Wahl der Primer und des Temperaturprofils dabei die größte Rolle. Die Spezifität der PCR läßt sich beispielsweise durch eine Erhöhung der Temperatur um 2°C - 5°C während der Anlagerung der Primer steigern. Nach Beendigung der Amplifikation wurden die Ansätze im Agarosegel oder PAGE aufgetrennt und anhand eines DNS-Längenstandards analysiert. Falls eine Weiterverarbeitung des PCR-Produktes erforderlich war, wurde die gewünschte Bande aus dem Gel präpariert (siehe S. 46). Generell wurde beim Ansetzen einer PCR versucht, so kontaminationsfrei wie möglich zu arbeiten, wobei insbesondere auf die Verwendung separater Arbeitsgeräte beim Pipettieren von PCR Ansätzen und Amplifikaten geachtet wurde. Als Primer wurden erprobte Standard-Primer verwendet oder neue mit dem Programm PRIMER ausgewählt, wobei besonderes Augenmerk darauf gelegt wurde, daß die Schmelzpunkte beider Primer möglichst nahe beieinander lagen, und daß die Primer nicht mehr als drei Basenpaare zu sich selbst oder zueinander komplementäre Sequenzen enthielten, so daß sich keine Primer-Dimere ausbilden konnten. Die Sequenz zwischen den Primern wurde auf Homologien untersucht, wobei eine möglichst geringe Anzahl erwünscht war. Die gelieferten Primer wurden für eine Stockkonzentration von 100 pmol/µl in Wasser gelöst und mit einer Konzentration von 5 pmol/µl aliquotiert.

3.15.2 PCR – Programme

Diese sind Tabelle 5 zu entnehmen.

Amplifikat	1. Denaturierung	Zyklen	Denaturierung	Anlagerung	Elongation	Extension	Endelongation
AKAP	2'/ 95°C	35	1'/ 95°C	1'/ 60°C	2'/ 72°C	keine	10'/ 72°C
AKAP (S)	2'/ 95°C	35	1'/ 95°C	1'/ 60°C	2'/ 72°C	5"	10'/ 72°C
SAGE - Ditags	10"/ 95°C	26	30"/ 95°C	1'/ 55°C	1'/ 70°C	keine	5'/ 70°C
Plasmide	20'/ 95°C	30	1'/ 95°C	1'/60°C	2'/ 72°C	keine	10'/ 72°C

Tabelle 5. PCR-Programme. Anmerkungen: *Amplifikat* - die zu amplifizierende DNS Matrize; *1. Denaturierung*- Initialer Denaturierungsschritt; *Zyklen* - Anzahl der Wiederholungen der Schritte Denaturierung, Anlagerung und Elongation; *Denaturierung* - Trennung des DNS Doppelstranges; *Anlagerung* Schritt, bei welchen sich die Primer anlagern; *Elongation* - Schritt, während dem die Polymerase den komplementären DNS Strang synthetisiert; *Extension* - die Zeit, um die der Elongationsschritt bei jedem Durchlauf verlängert wird; *Endelongation* - terminaler Elongationsschritt; *(S)* - Programme zur Fluoreszeinmarkierung der entsprechenden Sonden.

3.16 Fluoreszein – Markierung von DNS – Sonden

3.16.1 Markierung in der PCR

Für die Markierung von DNS-Sonden mittels PCR wurde ein Verfahren angewendet, das auf dem *Gene Images Kit* (Amersham) basiert. Das zu markierende Amplifikat einer vorangegangenen PCR wurde auf einem präparativen Agarosegel elektrophoretisch aufgetrennt, mit dem *QIAquick Gel Extraction Kit* gereinigt (siehe S.46) und nach erfolgter Optimierung der PCR in einem Reamplifizierungsschritt mit Fluoreszein markiert. Dies geschah im 50µl Ansatz in folgender Zusammensetzung:

- circa 10ng gelgereinigtes PCR-Amplifikat
- 5µl 10 x *Dynazyme*-Puffer
- je 1µl dATP, dGTP und dCTP (Konzentration: je 10mM)
- je 25pmol der beiden Primer
- 1µl dNTP (Konzentration: 1,2mM)
- 5µl Fluoreszein-dUTP (Konzentration: 1mM)
- 1,5µl *Dynazyme* (2U/µl)
- Ergänzung des Volumens mit Wasser

Die Überschichtung der Probe mit circa 30µl Öl, die Art und Weise des Pipettierens, der Warmstart und die Wahl des PCR Programms entsprach dem üblichen Vorgehen bei einer PCR (siehe S. 39).

3.16.2 Kontrolle des Fluoreszein-Einbaus

Der Erfolg der Fluoreszeinmarkierung der amplifizierten Sonden wurde in einem Standardagarosegel abgeschätzt. Nach der Gelelektrophorese wurde auf dem UV-Schirm die Bandenintensität mit einer Positivkontrolle verglichen, bei welcher in einem sonst identischen PCR-Reaktionsansatz dTTP anstelle von Fluoreszein-dUTP eingesetzt worden war. Um die Sonde zu reinigen, wurde diese aus dem Gel herausgeschnitten und per *QIAquick Gel Extraction Kit* (Qiagen) aufbereitet. Zur Abschätzung der Konzentration der gelgereinigten Sonde wurde ein Zehntel ihre Volumens im Agarosegel aufgetragen und die Menge anhand einer DNS-Leiter bekannter Konzentration bestimmt.

3.17 Northern – Blot

Die Durchführung der Northern-Blot Analyse erfolgte in Anlehnung an Sambrook (1989).

3.17.1 Transfer der RNS auf eine Nylonmembran

Denaturierung der RNS

Die RNS (10 - 40µg in 5,4µl) wurde mit 5,4µl deionisiertem Glyoxal, 16µl DMSO (Dimethylsulfoxid) und 3µl 0,1M Natriumphosphatpuffers (pH 7) versetzt und für eine Stunde bei 50°C inkubiert. Nach der Denaturierung wurden die Proben auf Eis gestellt und 4µl Probenpuffer (50% Glyzerin, 10mM Natriumphosphatpuffer (pH 7), 0,1% Bromphenolblau) hinzugegeben.

Gelelektrophoretische Auftrennung der RNS

Als Laufpuffer für das 1% Agarosegel wurde 1mM Natriumphosphatpuffer (pH 7) verwendet. Die Proben wurden mit einem Volumen von je 33,8µl aufgetragen. Zur Molekulargewichtbestimmung wurden ein DNS- und ein RNS- Größenstandard verwendet. Es wurde eine Spannung von 3 - 4 V/cm angelegt, was bei 100V eine Laufzeit von drei bis vier Stunden ergab. Dabei wurde der Laufpuffer kontinuierlich durch eine Umwälzpumpe umgewälzt. Nach einer Laufstrecke der Bromphenollaufbande von circa 11cm wurde das Gel auf dem UV-Schirm mit angelegtem fluoreszierendem Lineal fotografisch dokumentiert. Zur Orientierung wurde eine Ecke des Gels abgeschnitten.

Kapillartransfer der RNS

Die RNS wurde mit 20 x SSC als Puffer von dem Agarosegel auf eine Nylonmembran (Böhringer Mannheim) transferiert. Dazu wurde die Membran kurz in Aqua bidest. getaucht und für circa zehn Minuten in 2 x SSC geschüttelt. Mit fünf Lagen des *3MM*-Papiers (Whatman) wurde genauso verfahren. Der Transferturm wurde in folgender Reihenfolge aufgebaut: unten ein *3MM*-Papierstreifen, welcher an beiden Enden in das Pufferreservoir ragte, darauf das Gel, die feuchten *3MM*-Papiere, fünf trockene *3MM*-Papiere, mehrere Lagen saugfähiger Papiertücher und zuoberst ein Gewicht von circa 500 g. Die Zeit für den Transfer betrug zwölf bis fünfzehn Stunden. Nach Abbau des Turms und Markierung der Gelkanten auf der Membran wurde diese auf Gelgröße zurecht geschnitten und beschriftet. Anschließend wurde sie in 6 x SSC geschwenkt. Zum Binden der RNS an die Nylonmembran wurde diese für zwei Stunden bei 180°C inkubiert. Die Lagerung erfolgte trocken bei Raumtemperatur. Benutzte Membranen wurden feucht bei -20°C aufbewahrt.

3.17.2 Hybridisierung und Detektion von Northern-Blots mit dem *Gene-Images*-System

Hybridisierung

Die Hybridisierung mit fluoreszeinmarkierten DNS-Sonden erfolgte entsprechend dem mitgelieferten Protokoll des Herstellers (*Gene-Images*-System, Amersham). Die Prähybridisierung wurde bei 60 - 66 °C für mindestens zwei Stunden in einem Puffer mit 5% Dextransulfat, 5 x SSC, 0,1% SDS und 1/20 Volumenanteil *Liquid Block*-Reagenz (Amersham) durchgeführt. Die markierte Sonde wurde durch fünfminütige Inkubation bei 100°C und für weitere fünf Minuten auf Eis denaturiert und anschließend unter Schütteln zur die Prähybridisierungslösung gegeben. Pro Blot wurden 300 - 400ng Sonde verwendet. Die Hybridisierung erfolgte über Nacht bei 60 - 66°C unter ständigem Schütteln. Im Anschluß wurde die Membran kurz in einer Mischung aus 0,5 x SSC und 0,1% SDS gewaschen und dann bei 60 - 64°C für je 15 Minuten zweimal in der gleichen Lösung und in einem dritten Schritt in 0,1 x SCC und 0,1 SDS geschüttelt.

Detektion

Diese erfolgte mit einem Anti-Fluoreszein-Antikörper. Dafür wurde nach der Hybridisierung die Membran mit einem Verdünnungspuffer (100mM Tris-HCl [pH 7,5] und 300mM NaCl) gespült und für eine Stunde bei Raumtemperatur mit *Liquid Block*-Reagenz (1:10 in dem

Verdünnungspuffer) zur Blockierung unspezifischer Verbindungen inkubiert. Nach erneutem Spülen der Membran mit dem Verdünnungspuffer wurde sie eine Stunde lang bei Raumtemperatur mit Anti-Floureszein-Ig-Alkalische-Phosphatase-Konjugat (1:5000 in dem Verdünnungspuffer inklusive 0,5% BSA) inkubiert. Es schloß sich ein dreimaliger Waschvorgang von je 15 Minuten mit dem Verdünnungspuffer, welcher zusätzlich 0,3% Tween enthielt, an. Vor der eigentlichen Detektion wurde ein letztes Mal mit dem puren Verdünnungspuffer gespült. Die gebundenen Antikörperkonjugate wurden so dann mit einem Substrat für die alkalische Phosphatase detektiert, welches ein fluoreszierendes Produkt ergab. Dieses wurde über die Belichtung eines Films (Kodak BioMax MS) sichtbar gemacht, wobei die Expositionszeit zwischen einigen Sekunden und mehreren Tagen betrug.

Entfernen einer Sonde von einer Membran

Um nach einer Hybridisierung die Membran wiederverwenden zu können, mußte die Sonde entfernt werden. Hierzu wurde ein "Stripping"-Puffer (10mM Tris [pH 8], 1mM EDTA und 1% SDS) in einem großen Becherglas auf dem Bunsenbrenner zum Kochen gebracht. Bei einer Temperatur von 90 - 100°C wurde die Membran zehn bis fünfzehn Minuten inkubiert. Dieser Vorgang wurde mit einem frischen Puffer wiederholt. Bis die Membran erneut hybridisiert wurde, wurde sie bei -20°C feucht aufbewahrt.

3.18 cDNS Southern - Blot

Die Durchführung erfolgte in Anlehnung an Sambrook (1998).

3.18.1 Transfer von cDNS auf eine Nylonmembran

Die elektrophoretische Auftrennung der cDNS erfolgte im 1% Agarosegel bei einer Spannung von 1V/cm mit 1 x TAE als Laufpuffer. Das Gel wurde auf dem UV-Schirm fotographiert und die rechte untere Ecke zur Orientierung abgeschnitten. Für den Kapillartransfer wurde wie beim Northern Blot eine positiv geladene Nylonmembran (Böhringer Mannheim) verwendet. Zur Denaturierung der doppelsträngigen DNS wurde das Gel vor dem Transfer 45 Minuten bei Raumtemperatur in 1,5M NaCl und 0,5M NaOH inkubiert und nachfolgend kurz in Wasser gespült. Zur Neutralisierung wurde bei Raumtemperatur in 1,5M NaCl und 0,5M TrisHCl (pH 7,5) für 15 Minuten und ein zweites Mal für 30 Minuten inkubiert. Nach diesen Inkubationsschritten wurde das Gel, einige Lagen *3MM*-Papier (Whatman) und die Nylonmembran mit 2 x SSC gespült sowie der Transferturm (siehe S.43) mit einem 20 x SSC Reservoir als Puffer aufgebaut. Im Anschluß an den Transfervorgang, der zwölf bis fünfzehn

Stunden dauerte, wurde der Turm abgebaut und die Membranorientierung gekennzeichnet. Die Membran wurde kurz in 6 x SSC und dann für zwei Stunden bei 80°C im Ofen inkubiert. Sie wurde trocken bei Raumtemperatur beziehungsweise nach erfolgter Hybridisierung feucht bei -20°C gelagert.

3.18.2 Hybridisierung und Detektion von cDNS Southern - Blots mit dem *Gene-Images*-System

Diese erfolgte nach demselben Schema wie beim Northern - Blot (vergleiche S. 44). Allerdings wurde die Hälfte der dort angegebenen Menge an Sonde benötigt.

3.19 Präparative Gelelektrophorese

3.19.1 QIAqick Gel Extraction Kit

Zur Isolierung von DNS-Fragmenten einer Länge von 100 bis 10000 Basenpaaren aus Agarosegelen kam der *QIAqick Gel Extraction Kit* (Qiagen) zur Anwendung. Nach der elektrophoretischen Auftrennung der DNS wurde der Bereich des Gels, welcher das gewünschte Fragment enthielt, unter UV-Kontrolle ausgeschnitten. Die Agarose wurde anschließend in drei Volumina des mitgelieferten QX1 Puffers für zehn Minuten bei 50°C unter intermittierendem Invertieren inkubiert. Nach Verflüssigung des Gels wurde ein Volumen Isopropanol zugegeben und die Mixtur in einer vom Hersteller mitgelieferten Zentrifugationssäule für eine Minute bei 10000 x g zentrifugiert. Dieser Vorgang wurde nach der Zugabe von 500µl QX1 Puffers auf die Säule wiederholt. Zum Waschen wurde die Säule für fünf Minuten mit 750µl des mitgelieferten PE Puffers inkubiert und anschließend zweimal für eine Minute bei 10000 x g Umdrehungen zentrifugiert. Nach den Zentrifugationsschritten wurden die durchgelaufenen Lösungen jeweils verworfen. Zur Elution der gebundenen DNS wurde die Säule mit 50µl Tris-HCl (pH 8,5) für eine Minute inkubiert und durch eine einminütige Zentrifugation (10000 x g) aus der Membran der Säule gelöst. Bis zum weitern Gebrauch wurde die gereinigte DNS bei -20°C aufbewahrt.

3.19.2 Präparation von DNS-Fragmenten nach Polyacrylamidgelelektrophorese

Hierzu wurde nach der Auftrennung der DNS Probe im PAGE das Zielfragment nach SYBR Green I Färbung (vergleiche S. 35) auf dem UV-Schirm herausgeschnitten und in ein 0,5ml Mikrozentrifugenröhrchen, in dessen Boden zuvor mit einer Nadel ein circa 0,5mm großes Loch gebohrt worden war, plaziert. Dieses 0,5 ml Röhrchen wurde vor einer zweiminütigen Zentrifugation bei 10000 x g und 4°C in ein 1,5ml Gefäß gesetzt. Auf diese Weise wurden die exzidierten Gelbanden fragmentiert in dem äußeren Gefäß gesammelt. Zu diesen Fragmenten

wurden jeweils 800µl LoTE dazugegeben und über Nacht bei 4°C inkubiert. Im nächsten Schritt wurde nach einer 15 minütigen Inkubation bei 65°C oder 37°C die Mixtur in *SpinX* (0,22µm)-Zentrifugationssäulen (Costar Inc.) geladen und für mindestens fünf Minuten bei 10000 x g und 4°C zentrifugiert, bis die Lösung durch die gesamte Säule gelaufen war. Dieser Durchlauf wurde in 300µl Portionen aufgeteilt und einer Ethanolfällung unterzogen. Anschließend wurden die gefällten und gewaschenen DNS Pellets in je 10µl LoTE resuspendiert und in einem Reaktionsgefäß gesammelt.

3.20 Binden von biotinylierten cDNS-Fragmenten an *Dynabeads*

Um mit Biotin markierte cDNS-Fragmente isolieren zu können, wurden sie an magnetische mit Straptavidin bedeckte Partikel (*Dynabeads*) gebunden.

Zur Vorbereitung wurden die Partikel (50µl, entspricht 0,5mg) mittels eines entsprechenden Magnets immobilisiert und der Überstand abgenommen. Zum Waschen wurden 200µl des B&W Puffers zugegeben, die Partikel aufgemischt, wieder immobilisiert und der Puffer entfernt. Um die Biotin-cDNS Fragmente zu binden, wurden 100µl B&W Puffer, 90µl LoTE und 10µl der cDNS-Lösung hinzugefügt. Anschließend wurde bei Raumtemperatur für zwanzig Minuten inkubiert und dreimal mit je 200µl B&W Puffer und einmal mit 200µl LoTE gewaschen.

3.21 Vektorisolation

3.21.1 Reinigen von Plasmid-DNS mit dem *Plasmid Mini Purifikation Kit* (Qiagen)

Die Präparation der DNS erfolgte entsprechend dem von Qiagen mitgelieferten Protokoll. Die über Nacht in 3ml LB-Medium gewachsenen *E. coli* Kulturen wurden halbiert. Nach Zentrifugation bei 5000 x g (4°C) für fünf Minuten und Entfernung des Überstandes wurden zum Resuspendieren des Bakterienpellets diese mit 300µl des mitgelieferten P1 Puffers versetzt. Um die Bakterien zu lysieren, wurde den Ansätzen Puffer P2 (300µl) zugegeben und bei Raumtemperatur fünf Minuten inkubiert. Nach Zugabe von 300µl P3 zur Präzipitation von Proteinen und genomischer DNS erfolgte eine Inkubation von zehn Minuten auf Eis. Nach Zentrifugation bei 15000 x g (4°C) für 15 Minuten wurde der Überstand, der die Plasmide enthielt, sofort entfernt und auf die mit 1ml QBT Puffer equilibrierten Reinigungssäulen gegeben. Um sämtliche kontaminierenden Bestandteile zu entfernen, wurde viermal mit je 1ml QL Puffer gewaschen. Zur Elution der DNS wurden 800µl QF Puffer auf die Säule pipettiert. Die gelöste DNS wurde mit je 560µl Isopropanol gefällt, zweimal mit Ethanol gewaschen und in je 20µl LoTE aufgenommen.

3.21.2 Reinigen von Plasmid-DNS mit *Qiaprep 96 Turbo Miniprep Kit*

Dieser Schritt sowie die Sequenzierung mit dem *BigDye™ Terminator Cycle Sequencing Kits* (Perkin Elmer, Vaterstetten, BRD) wurde in Kooperation mit dem Institut für Molekulare Biotechnologie (IBM) in Jena durchgeführt.

Die Präparation der DNS erfolgte entsprechend dem von Qiagen mitgelieferten Protokoll. Die über Nacht in LB-Medium gewachsenen *E. coli* Kulturen wurden zum Resuspendieren des Bakterienpellets mit 250µl des mitgelieferten P1 Puffers versetzt. Nach Transfer der Suspension in einen speziellen Block, welcher Teil des Kits ist, wurde jeder Probe 250µl des Puffers P2 zugesetzt, der Block mehrfach invertiert und bei Raumtemperatur für fünf Minuten inkubiert. Nach anschließender Zugabe von 350µl Puffer N3 wurden die Zellysate in die *TurboFilter* Platten des Herstellers pipettiert. Mit einer Geschwindigkeit von ein bis zwei Tropfen pro Sekunde wurden durch Anlegen eines Vakuums denaturierte und gefällte Zellbestandteil abgefiltert, während die Partikel freien Lysate im QIAprep Modul aufgefangen wurden. Die an eine Silicagelmembran gebundene DNS wurde durch erneutes Anlegen eines Vakuums von RNS, zellulären Proteinen und Metaboliten gereinigt. Nach zweimaligem Waschen mit 0,9ml PE Puffer wurden die Membranen durch zehnminütiges Anlegen des maximalen Vakuums getrocknet. Zur Elution der DNS wurden 100µl EB Puffer (10mM TrisHCl mit einem pH von 8,5) verwendet.

3.22 Sequenzierung

3.22.1 Sequenzierung mit fluoreszierenden Primern

Nach Reinigung einer Über-Nacht-Kultur des Plasmids (siehe S.47) erfolgte die Sequenzierung anhand des *Thermo Sequenase fluorescent labelled primer cycle sequencing kit* (Amersham, Braunschweig, Deutschland) und eines *ALFexpress DNA* Sequenzierautomaten (Pharmacia Biotech, Freiburg, Deutschland). Dazu wurden zu 5µl des Plasmids 1µl fluoreszierender Primer (M13 universal; Konzentration: 1,5pmol/µl) und 2µl der vier Nukleotidgemische gegeben, mit 30µl Öl beschichtet und folgendem Programm unterzogen: 95°C für 30 Sekunden, 20 Zyklen mit 95°C für 30 Sekunden und 72°C für 1 Minute, abschließend 72°C für 5 Minuten. Nach Entfernung des Öls wurden je 5µl Ladepuffer zugegeben und das Gemisch auf das 6% Polyacrylamid-Gel (15ml 30%Acrylamidlösung, 7ml Aqua bidest., 37,5ml 47% Harnstofflösung, 15ml 5 x TBE-Puffer, 375µl 10% APS und 30µl TEMED) aufgetragen.

3.22.2 Sequenzierung mit fluoreszierenden Nukleotiden

Die zu sequenzierende Plasmid-DNS wurde vor der Sequenzierung mittels *Qiaprep 96 Turbo Miniprep Kit* (Qiagen) (siehe S.48) gereinigt. Die Sequenzierung erfolgte unter Verwendung eines ABI PRISM® *DNA Sequencing Kits* (Perkin Elmer, Vaterstetten, BRD), nämlich des *BigDye™ Terminator Cycle Sequencing Kits*, und von DNS Sequenzierautomaten (ABI 373A/ ABI Perkin Elmer, Weiterstadt, BRD), wobei die Terminatoren - Didesoxynukleotide - markiert waren. Entsprechend den Empfehlungen des Herstellers wurde nach Zugabe von pro Sequenzierungsansatz 4µl Kit (enthielt Puffer, AmpliTaq FS DNA Polymerase, dNTPs und die Terminatoren), 3µl H_2O, 1µl des M13 Vorwärtsprimers zu 4µl der aufbereiteten DNS-Lösungen die Sequenzierungs-PCR mit folgendem Programm durchgeführt: 3 Minuten: 95°C, 10 Sekunden: 95°C, 1 Minute: 50°C, 4 Minuten: 60°C. Die Schritte zwei bis vier wurden 34 mal wiederholt, bevor die Reaktion auf 4°C gekühlt wurde. Um nicht eingebaute markierte ddNTPs zu entfernen, wurde ein Ethanolpräzipitation durchgeführt. Die gefällte DNS wurde in 2µl Probenpuffer (1ml Ansatz: 40µl EDTA [0,5 M], 960µl Formamid, 6mg Dextranblau) aufgenommen und für fünf Minuten bei 80°C denaturiert, bevor sie auf das Gel aufgetragen wurde. Für die Gelelektrophorese und zur Detektion wurde der oben genannte Sequenzierautomat benutzt.

3.23 Statistische Methoden

Im Folgenden wird eine Test übergreifende Berechnung dargestellt (Berechnung der Regulation). Die Beschreibung sämtlicher weiterer Methoden erfolgt im Rahmen der Darstellung des jeweiligen Testes.

Um das Ausmaß des Unterschieds zwischen den beiden Kontrollprofilen berechnen zu können, wurden - um Division durch Null zu vermeiden - sämtliche Nullen durch den Wert eins ersetzt (nur für diese Berechnung). Anschließend wurden die so entstandenen Gesamthäufigkeiten berechnet (Daten ohne Sequenzfehlerkorrektur beispielsweise: K1 von 13584 auf 19462 und K2 von 13915 auf 20036 gestiegen) und zur Normalisierung (Lal et al. 1999) K1 auf K2 hochgerechnet (Daten ohne Sequenzfehlerkorrektur: Faktor 1,0295). Die Division von K1 durch K2 ergab die Faktoren der "Regulation". Im Rahmen der vorliegenden Arbeit werden als "reguliert" nur die Tagpaare betrachtet, deren Faktor $\leq 0{,}5$ beziehungsweise ≥ 2 war.

4 Etablierung von SAGE

4.1 Ergebnisse der Etablierung

4.1.1 Ausgangssituation und Ziel

Um Fragen der Genexpression zu etablierten Tiermodellen wie zum Beispiel Schlaganfall umfassender als zuvor beantworten zu können, wurde angestrebt, eine entsprechende Methode zu etablieren. Aus den eingangs dargestellten Gründen wurde SAGE gewählt. Das ermittelte Expressionsprofil gesunder Mäusegroßhirne sollte später als Vergleichswert in Studien zur Genexpression pathologischer Zustände dienen. Ziel des folgenden Kapitels ist es, die Resultate der Etablierung dieser Methode darzustellen.

4.1.2 Zusammenfassende Beschreibung der Methode

Die serielle Analyse der Genexpression wurde anhand des von Velculescu et al. (1997b) in der Version 1.0c freundlicher Weise zur Verfügung gestellten Protokolls durchgeführt.
Zur Herstellung einer SAGE-Bibliothek wurde als erstes extrahierte Boten-RNS in doppelsträngige cDNS umgewandelt (siehe Abbildung 3). Hierbei wurden während der Erststrangsynthese Oligo(dT)Primer verwendet, die am 5' Ende biotinyliert waren. Auf diese Weise konnte die cDNS am 3' Ende des ursprünglichen Transkriptes später über die Bindung an Streptavidin erfaßt werden. Die cDNS wurde zunächst mit dem Restriktionsenzym NlaIII verdaut. Diese im Kontext von SAGE Verankerungsenzym genannte Endonuklease erkennt und schneidet die DNS unmittelbar 3' der Sequenz CATG (2). Dieser Schritt generierte eine definierte Lokalisation innerhalb des Transkriptes, da über die Biotinylierung dasjenige Fragment gekennzeichnet war, das dem 3' Ende des Transkriptes am nächsten liegt. Die biotinylierten cDNS Stücke wurden unter Verwendung von Streptavidin bedeckten magnetischen Partikeln affinitäts-gereinigt. Die immobilisierten cDNS Fragmente wurden in zwei Gruppen geteilt und jeweils am 5' Ende mit den Linkern A beziehungsweise B ligiert (3). Diese Linker enthalten einen der NlaIII Schnittstelle entsprechenden vier Basen langen Überhang, eine Erkennungssequenz für BsmFI (Typ IIS Restriktionsenzym) und eine Anlagerungssequenz (A beziehungsweise B) für PCR Primer. Die so verlängerten cDNS Stücke wurden mit BsmFI verdaut, welches 14 (20%) bis 15 (80%) (Madden et al. 2000) Basen 3' seiner Erkennungssequenz schneidet (4). "Tag" meint im Zusammenhang mit SAGE also die Nukleotidabfolge unmittelbar 3' der NlaIII Erkennungssequenz eines Transkriptes, die dem Poly-A-Schwanz am nächsten liegt. Die auf diese Weise von ihrer Verankerung an

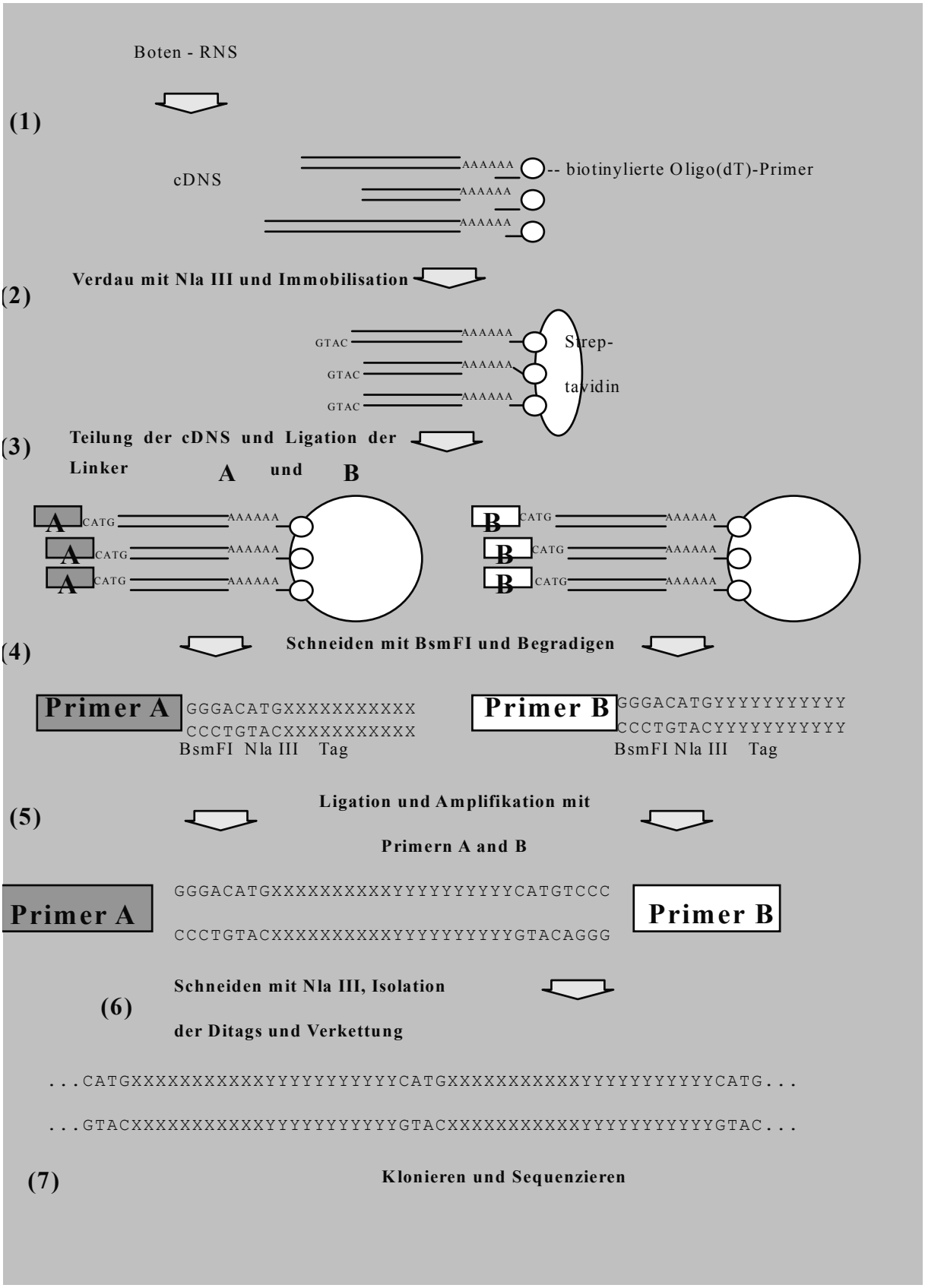

Abb. 3: Schema eines SAGE Durchlaufes.

die magnetischen Streptavidinpartikel abgeschnittenen und mit den beiden Linkern

verbundenen, ungefähr 11bp langen, SAGE Tags wurden mittels Klenow DNS Polymerase begradigt **(4)**.

Die beiden Gruppen wurden wieder vereinigt und mit T4 Ligase ligiert **(5)**. Die daraus resultierenden Ditags, welche von den beiden Linkern eingerahmt werden, wurden unter Verwendung der Primer A und B amplifiziert **(5)** und anschließend mit NlaIII verdaut, um die Linkersequenzen wieder zu lösen **(6)**. Abschließend wurden die Ditags mittels T4 Polymerase zu langen Ketten verbunden, die in einen Vektor kloniert und sequenziert wurden **(7)**, wobei die Erkennungssequenz von NlaIII als Interpunktion zwischen den Ditags dient.

4.1.3 Vorbereitende Tests

Vor dem Beginn eines SAGE-Durchlaufes sind Versuche notwendig, um den Grad der Biotinylierung der Oligo(dT)-Primer sowie der Kinasierung der Linker A und B zu überprüfen.

4.1.3.1 Streptavidin Gelshift - Assay

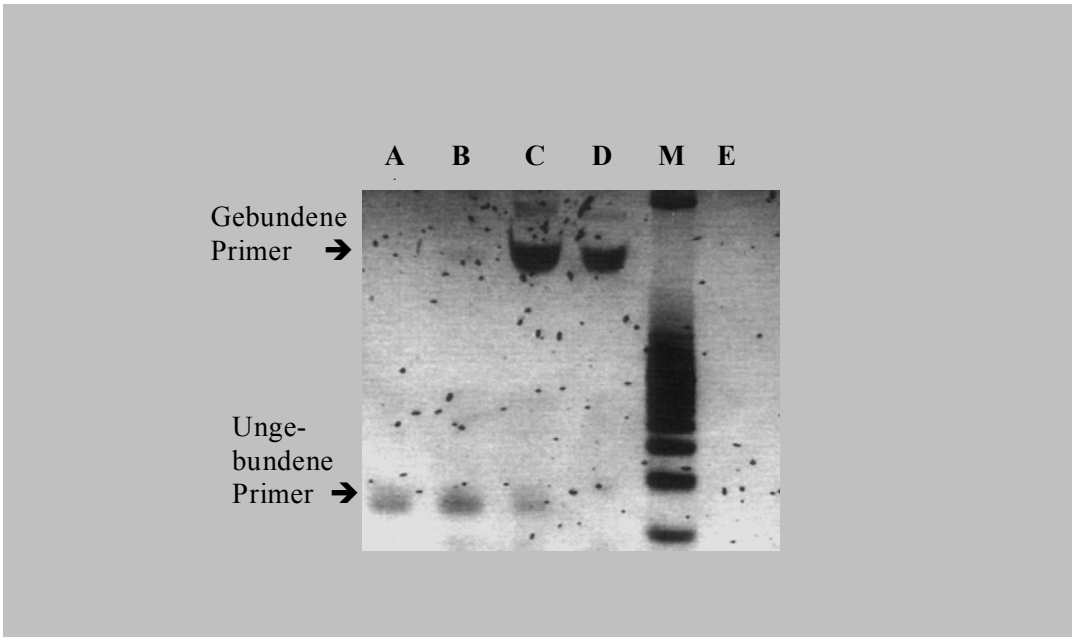

Abb. 4: Oligo (dT)-Biotin-Shift (12% PAGE). A: Primer (460ng), **B:** die gleiche Menge Primer mit 0,2µg Streptavidin, **C:** mit 1µg Streptavidin, **D:** mit 2µg Streptavidin, **E:** 0,4µg Streptavidin pur, **M:** 1kb DNS-Größenstandard.

Eine optimale Biotinylierung der Oligo(dT)-Primer, welche zur cDNS Synthese verwendet werden, stellt eine wesentliche Voraussetzung für eine effiziente Bindung der 3' Enden der cDNS an die magnetischen Partikel dar. Zum Überprüfen des Biotinylierungsgrades der

Oligo(dT)-Biotin Primer wurde ein Streptavidin Gelshift Assay in einem 12% Polyacrylamidgel durchgeführt.

Die HPLC gereinigten Oligo(dT)-Biotin Primer (Eurogentec) wurden für eine Stockkonzentration von 100pmol/µl in LoTE gelöst. Für den Assay wurden folgende 10µl Ansätze, deren Volumen entsprechend mit LoTE ergänzt worden war, für zwei Stunden bei Raumtemperatur inkubiert:

A 0,7µl Primer pur (entspricht 460ng)

B 0,7µl Primer + 1µl Streptavidin (entspricht 0,2µg)

C 0,7µl Primer + 5µl Streptavidin (entspricht 1µg)

D 0,7µl Primer + 10µl Streptavidin (entspricht 2µg)

E 2µl Streptavidin pur (entspricht 0,4µg)

Wie Abbildung 4 zeigt, nahm die Anzahl der ungebundenen Primer mit steigender Zugabe von Streptavidin kontinuierlich ab. Bei einem Überangebot an Streptavidin (Ansatz D) wurden die Primer allesamt gebunden. Das bedeutet, daß der Anteil nicht biotinylierter Primer vernachlässigbar war.

4.1.3.2 Selbstligation der Linker

Um die Kinasierung der Linker zu überprüfen, wurden diese im Vorfeld einer Selbstligation unterzogen, wobei davon ausgegangen wird, daß nicht phosphorylierte Oligomere sich nicht ligieren lassen. Ein geringer Phosphorylierungsgrad hätte im eigentlichen SAGE-Durchlauf aufgrund einer ineffizienten Ligation der Linker an die SAGE Tags einen Substanzverlust zur Folge.

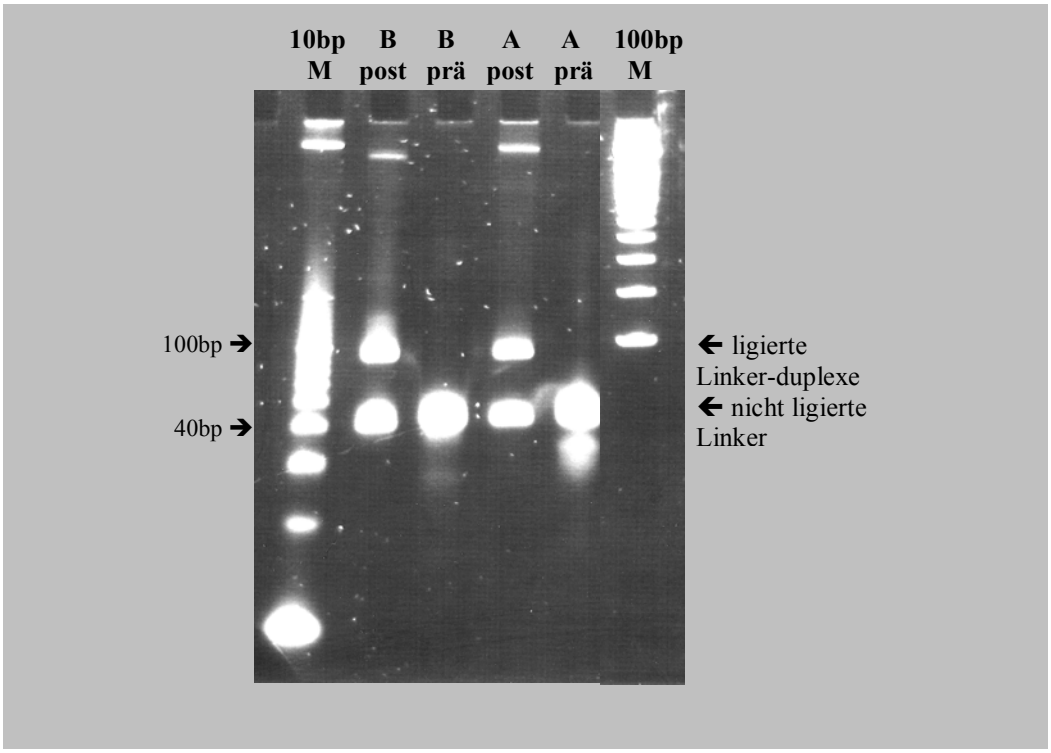

Abb. 5: Selbstligationstest der Linkerduplexe. 12% PAGE mit je 6µl Linkerduplexlösungen. A/B prä: die aneinander gelagerten Linkerduplexe A oder B vor der Ligation (je 44bp lang), post: nach der Selbstligation (Dimere: je 84bp lang). Die Selbstligationsrate liegt bei circa 50%. 10/100bp M: DNS Größenstandards.

Duplexbildung

Bevor die Verwendung der Linker jedoch möglich war, mußten die PAGE gereinigten einzelnen Linkerstränge (Eurogentec) zuerst in einem 100µl Ansatz mit einer Konzentration von je 100ng/µl anhand des folgenden Programms aneinandergelagert werden: 95°C für zwei Minuten, 65°C für zehn Minuten, 37°C für zehn Minuten, 19°C für zwanzig Minuten. Hierbei gewährleistet der erste Schritt die Denaturierung der Linker und die folgenden drei die Formierung von Duplexen. Durch das abfallende Temperaturprofil wird zu Beginn eine spezifische Anlagerung der beiden Stränge gesichert, um dann bei niedrigeren Temperaturen den Prozeß zu vervollständigen. Bis zum weiteren Gebrauch wurden die Duplexe bei -20°C aufbewahrt.

Ligationstest

Im 15µl Ansatz wurde die Selbstligation der Linkerduplexe folgendermaßen getestet: Eine Mischung von 3µl 5x Ligasepuffer und 11µl Duplex-Linkerlösung in einer Konzentration von 100ng/µl wurde zwei Minuten lang bei 50°C und für 15 Minuten bei Raumtemperatur inkubiert, um die Überhänge der Linker zu linearisieren, ohne durch zu hohe Temperaturen den gebildeten Doppelstrang wieder aufzutrennen. Anschließend wurde 1µl T4 Ligase (5

Units/µl, GibcoBRL) zugegeben und für zwei Stunden bei 16°C inkubiert. Nach Phenol-Chloroform-Extraktion und Ethanolfällung wurden die Ansätze in 6µl LoTE gelöst und in einem 12% Polyacrylamidgel elektrophoretisch aufgetrennt (Abb. 5). Obwohl nur eine Ligationsrate von schätzungsweise 50% resultierte - im Protokoll von Velculescu et al. (1997b) werden mindestens 70% Selbstligationsrate gefordert, wurden die Linker aus finanziellen Überlegungen weiter verwendet und ein entsprechender Effizienzverlust in Kauf genommen.

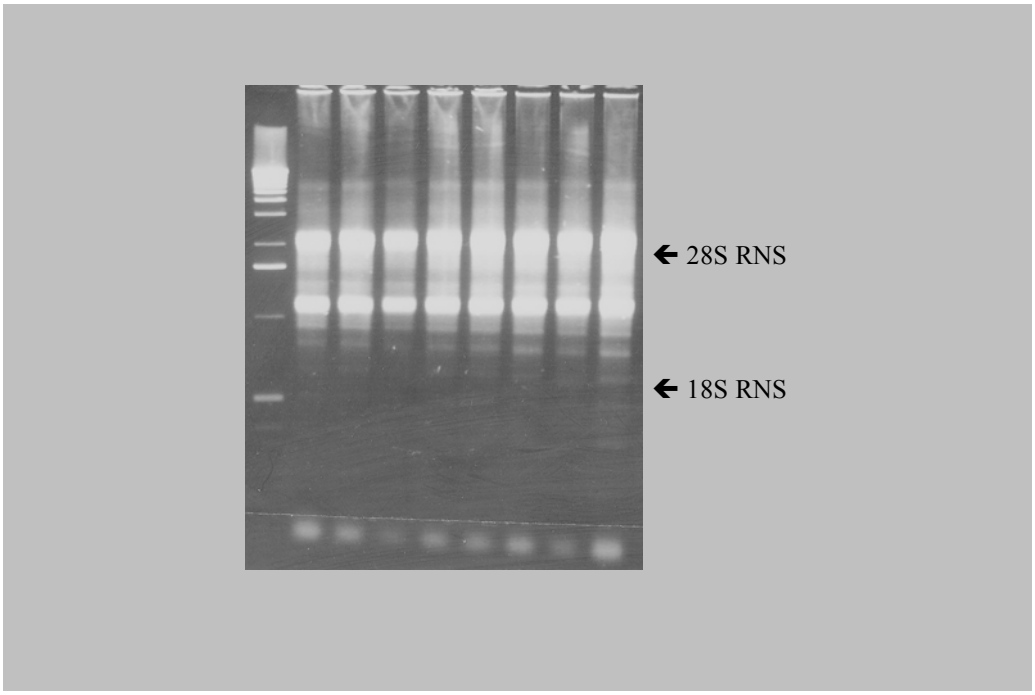

Abb. 6: Gesamt-RNS. Auftrennung der Kontroll-RNS (je 2µl) im 1% Agarosegel.

4.1.4 SAGE-Durchlauf

Nach dem erfolgreichen Abschluß der beiden Vorteste konnte mit dem eigentlichen Durchlauf von SAGE begonnen werden.

4.1.4.1 Isolierung der Gesamt-RNS

Aus den Großhirnhemisphären von vier erwachsen männlichen Mäusen desselben Stamms (C57b6, BGVV Berlin) mit einem mittleren Körpergewicht von 20 ± 2g wurde die Gesamt-RNS wie im methodischen Abschnitt dargestellt (vergleiche S. 37) extrahiert, vereinigt, gemischt und in zwei gleich große Volumina (jeweils 800µl) aufgeteilt (K1 und K2). Zur Überprüfung der Reliabilität wurden diese beiden Gruppen getrennt, jedoch parallel behandelt.

Die Messung der optischen Dichte ergab eine Konzentration der Gesamt-RNS von circa 0,8µg/µl, so daß bei einer Weiterverarbeitung von je 500µl Gesamt-RNS-Lösung eine Menge von je 10µg Boten-RNS zu erwarten war.[4] Das Verhältnis der OD_{260}/OD_{280} von größer zwei in beiden Fällen sprach für die Reinheit der extrahierten RNS. Abb. 6 zeigt in der Standardelektrophorese zwei diskrete Banden für die 18S und 28S ribosomale RNS, die in einem Verhältnis von 1 zu 1,5 - 2 zueinander stehen, so daß auf eine erfolgreiche Präparation ohne Degradierung der RNS geschlossen werden konnte.

4.1.4.2 Präparation der Boten-RNS aus der Gesamt-RNS

Die Präparation der Boten-RNS aus der Gesamt-RNS gestaltete sich wie beschrieben (siehe S. 38). Hier ergab die anschließende Messung der optischen Dichte eine Konzentration von 640ng/µl für K1 und 660ng/µl für K2. Bei einem Volumen von 20µl entspricht dies einer Gesamtmenge von 12,8µg (K1) beziehungsweise 13,2 (K2). Als Ausgangsmaterial für die cDNS Synthese lieferten im Falle von K1 7,8µl die erforderlichen 5µg Boten-RNS, im Falle von K2 waren dies 7,5µl. 3,1µl (K1) beziehungsweise 3,0µl (K2) wurden zur Analyse im Northern Blot verwendet.

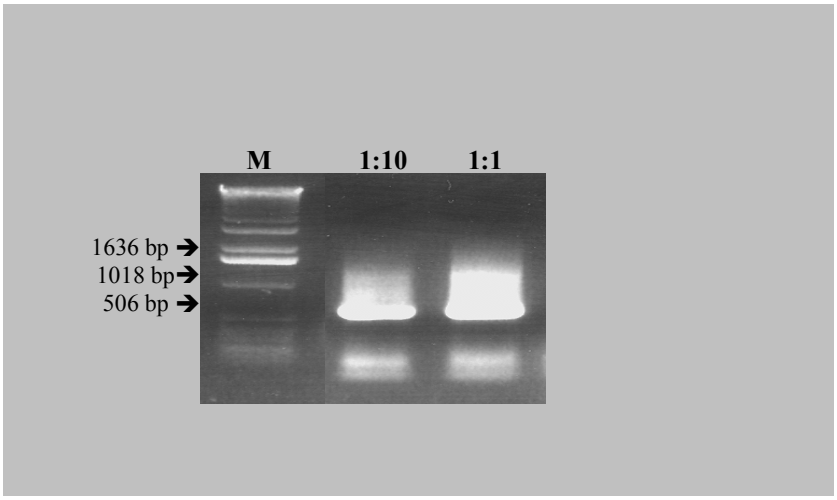

Abb. 7: Reamplifikations-PCR Akap149 (B5). B5IAT7 Rev - IIB5EF (670 bp). Um die PCR Bedingungen zu optimieren, wurden jeweils 1 µl des gereinigten PCR Produktes pur beziehungsweise aus einer 1:10 Verdünnung als Matratze eingesetzt. PCR-Programm: 95°C/1', 35 Zyklen mit 95°C/1', 60°C/2', 72°C/2' mit 5" Extension der Elongationszeit bei jedem Zyklus, abschließend 72°C/10'. Die 1:10 Verdünnung erwies sich als einerseits spezifisch amplifizierbar, andererseits war die produzierte Menge für eine Sonde ausreichend genug, so daß diese PCR als Grundlage für die folgende Markierungs-PCR verwendet werden konnte. M: 1kb DNS Größenstandard.

[4] Nach Angabe des Herstellers (GibcoBRL) liefert das zur Isolation von Boten-RNS verwendete System mindestens eine Ausbeute von 2%.

4.1.4.3 Northern - Blot

Um die Güte der Boten-RNS überprüfen zu können, wurde zu diesem Zeitpunkt mit den oben genannten Volumina der Boten-RNS-Lösungen (je 2µg) ein Northern Blot durchgeführt. Die Herstellung der Membran erfolgte wie dargestellt (siehe S. 44).

Als Sonde wurde ein Fragment des Proteinkinase A - Ankerproteins 149 (Akap149) verwendet, da dessen Transkriptvarianten eine Länge von mehreren Kilobasen aufweisen und so gut zur Qualitätsüberprüfung von Boten-RNS geeignet sind. Da die zur Amplifikation benutzten Primer der humanen Variante entsprechen, wurde die Sonde unter Verwendung muriner cDNS (vergleiche S. 38) und der Primer B5IAT7 Rev und IIB5EF wie beschrieben (siehe S. 42) hergestellt (siehe Abb. 7).

Da die Primer zuvor noch nicht in Versuchen mit Mäusen eingesetzt worden waren, wurde die Identität der Sonde anhand eines Restriktionsenzymverdau überprüft. Die Fragmente wurden mit per Internet erstellten Sequenzanalysen (www.firstmarket.com/cutter/cut2.html) verglichen. Die Auftrennung der geschnittenen Fragmente (keine Abbildung) zeigte eine gute Übereinstimmung mit der theoretischen Vorhersage, so daß die Spezifität der Sonde als erwiesen angenommen werden konnte.

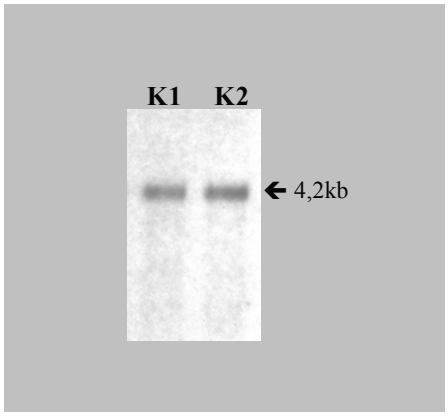

Abb. 8: Northern Blot mit Akap149. Es wurden pro Kontrollgruppe (K1, K2) 2µg Boten-RNS aufgetragen.

Ergebnis der Hybridisierung

Das bereits in diversen humanen Geweben und im Gehirn der Ratte detektierte 4,2kb große Fragment war auch in der Maus zu sehen (siehe Abb. 8). Diese Detektion des nicht degradierten Transkriptes ließ den Schluß zu, daß die Qualität der RNS den Erfordernissen von SAGE entsprach und weiter verwendet werden konnte.

4.1.4.4 cDNS Herstellung mit Oligo(dT)$_{20}$-Biotin Primern

Die Weiterverarbeitung der zurückbehaltenen Boten-RNS erfolgte wie dargestellt (siehe S. 38 im Methodenteil). Anstelle der vom Hersteller des cDNS Synthese Kits (GibcoBRL) mitgelieferten Oligo(dT)-Primer wurden jedoch jeweils 2,5µg der 5'-biotinylierten Oligo(dT)$_{20}$ Primer verwendet. Diese waren zuvor in einem Streptavidin-Gelshift-Assay auf ihren Biotinylierungsgrad getestet worden. Die gewonnene cDNS wurde in 21µl LoTE aufgenommen. 1/10 dieser cDNS wurde zur Analyse im 1% Agarosegel und anschließendem cDNS Southern Blot verwendet. Wie Abbildung 9 zeigt, erzielte die cDNS Synthese Fragmente, welche einen Längenumfang von mehreren hundert Basen bis zu über 10 Kilobasen aufwiesen und damit den gesamten Bereich der Boten-RNS abdeckten.

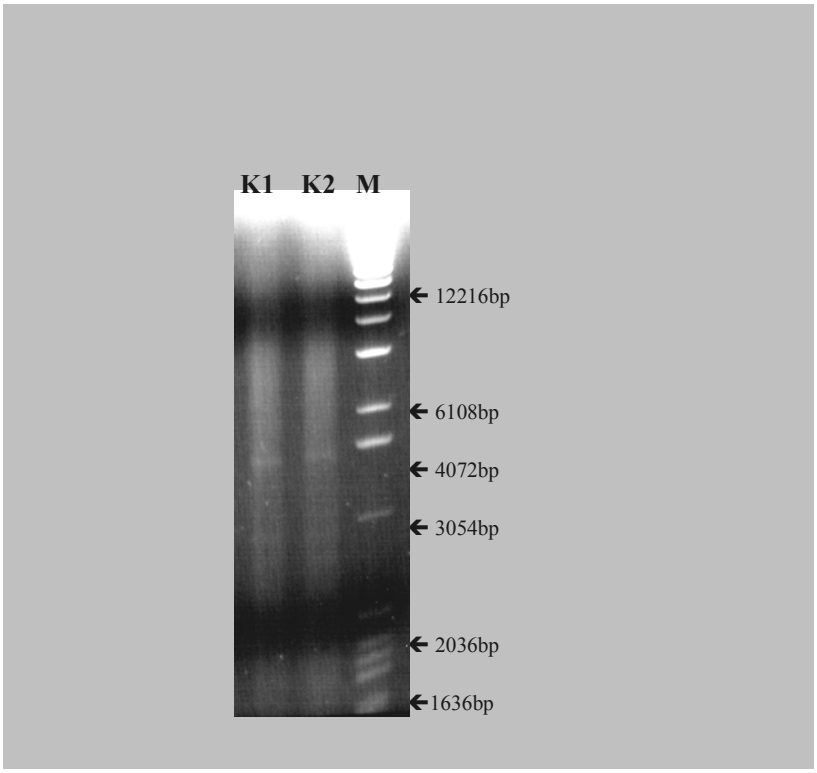

Abb. 9: cDNS Auftrennung. 1% Agarosegel. Diese Elektrophorese ergibt eine Länge der cDNS Fragmente von ungefähr 200bp bis größer als 12kb. M: 1kb DNS-Größenstandard, K1 und K2: 1,9µl cDNS (von je 21µl).

4.1.4.5 cDNS Southern Blot

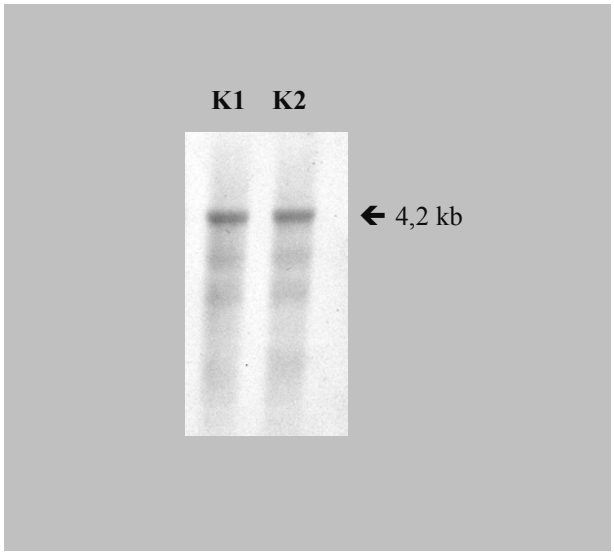

Abb. 10: cDNS Southern Blot mit Akap149. Die Menge an cDNS entspricht jeweils einem Zehntel der aus je 5µg Boten-RNS synthetisierten cDNS. K1 und K2: die beiden Kontrollgruppen.

Dieser wurde wie beschrieben durchgeführt (siehe S.45). Als Sonde wurde die für den Northern Blot erstellte Sonde (Akap149) verwendet. Das Ergebnis der Hybridisierung (siehe Abb. 10) zeigt ein höheres Hintergrundsignal als dasjenige des Northern Blots, was darauf zurückzuführen sein könnte, daß die Bindung zwischen zwei DNS Stücken weniger stabil ist, als diejenige zwischen RNS und DNS wie es beim Northern Blot der Fall ist (Sambrook et al. 1998²). Dennoch konnte auch hier das 4,2kb lange Transkript ohne Hinweis auf Degradierung als singuläre Bande detektiert werden, so daß mit dem SAGE-Durchlauf fortgefahren werden konnte.

4.1.4.6 Restriktionsenzymverdau der biotinylierten cDNS mit dem Verankerungsenzym NlaIII

Um aus den die gesamte Boten-RNS Länge umfassenden cDNS Stücken die SAGE Tags herzustellen, wurde die biotinylierte cDNS mit NlaIII (NEB) geschnitten.

Hierzu wurde die Hälfte (10µl) der synthetisierten cDNS eingesetzt (vergleiche S. 35). Die andere Hälfte der Proben wurde für spätere weiterführende Untersuchungen zurückbehalten. Es wurden pro Ansatz 50 Units einer frisch gelieferten Charge NlaIII verwendet. Nach einstündiger Inkubation bei 37°C, Extraktion und Ethanolfällung (S.33) wurde die geschnittene cDNS in 20µl LoTE aufgenommen.

4.1.4.7 Binden der biotinylierten cDNS-Fragmente an magnetische mit Streptavidin bedeckte Partikel

Im nächsten Schritt sollten diejenigen Fragmente der cDNS, welche an Biotin gebunden waren, vom Rest der cDNS getrennt werden. Auf diese Weise wurden nur diejenigen cDNS Fragmente weiterverarbeitet, die das am meisten 3' liegende NlaIII Fragment jedes Transkriptes repräsentierten und so einem definierten Ort der Boten-RNS entstammten. Hierzu wurden pro Kontrollgruppe je zweimal 100µl *Dynabead* M-280 Streptavidin Suspension (Dynal) eingesetzt, was jeweils einer Menge von 1mg entsprach, und die Hälfte der verdauten cDNS (vergleiche S. 47), also jeweils 20µl in zwei Ansätzen. In diesem Schritt wurde die cDNS der beiden Kontrollgruppen jeweils in die Untergruppen A und B aufgeteilt, um sie im darauffolgenden Schritt entweder mit Linker A oder mit Linker B ligieren zu können.

4.1.4.8 Ligation der Linker an die gebundene cDNS

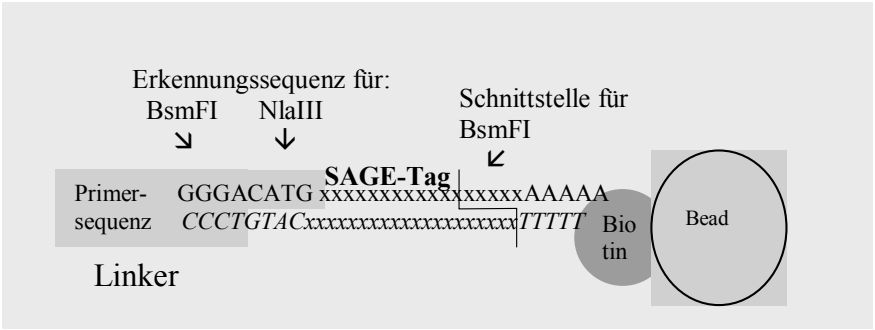

Abb. 11: Die Entstehung eines SAGE-Tags.

Es folgte die Ligation der Linker A und B an die in die Untergruppen A und B geteilte und an die *Dynabeads* gebundene cDNS-Fragmente. Die Oligomerduplexe (Linker) enthalten eine Primerbindungssequenz (A beziehungsweise B), die Erkennungssequenz für das später verwendete '*tagging*' Enzym BsmFI und einen der Schnittstelle von NlaIII entsprechenden Überhang (siehe Abb. 11). Dieser Schritt des SAGE Protokolls ermöglicht später das Herausschneiden des SAGE Tags durch das Typ IIS Restriktionsenzym BsmFI und die PCR-Amplifikation der Ditags. Der Zustand der Kinasierung der Linkerduplexe war im Vorfeld durch Selbstligation untersucht worden (siehe S. 53). Die Ligation der Linkerduplexe A und B wurde sofort im Anschluß an das Binden der cDNS Fragmente an die *Dynabeads* durchgeführt. Hierzu wurde jeweils die gesamte Menge der immobilisierten cDNS und 2µg von einem der beiden Linkerduplexe eingesetzt und wie auf S. 36 beschrieben verfahren.

4.1.4.9 Restriktionsenzymverdau der gebundenen cDNS mit dem Tags produzierendem Enzym BsmFI

Die Typ IIS Restriktionsendonuklease BsmFI schneidet in 20% der Fälle 14 Basen und in 80% der Fälle 15 Basen (Madden 2000) 3' ihrer Erkennungssequenz asymmetrisch (siehe Abb. 11). Durch den Verdau mit diesem Enzym wurden die SAGE Tags mitsamt den 5' ligierten Linkern A oder B von dem 3'-terminalen cDNS-Anteil abgeschnitten, welcher an den magnetischen Partikeln gebunden blieb. Hierzu wurden pro Ansatz 4U BsmFI (NEB) eingesetzt und eine Stunde bei 65°C inkubiert. Anschließend wurden die abgeschnittenen Transkriptreste samt PolyA-Schwanz magnetisch immobilisiert und der die Tags enthaltende Überstand abgenommen. Nach Phenol-Chloroform-Extraktion und hochkonzentrierter Ethanol-Fällung wurden die vier Ansätze in je 10µl LoTE aufgenommen.

4.1.4.10 Begradigung der Tags

Wie der Abbildung 11 zu entnehmen ist, wiesen die Linker-Tag-Komplexe 5' am komplementärer Strang einen Überhang auf, der durch den Verdau mit BsmFI verursacht worden war. Um die Ligation der Tags der beiden Untergruppen A und B miteinander zu ermöglichen, mußten diese Enden begradigt werden. Dazu wurde das Enzym Klenow wie auf S. 36 beschrieben verwendet. Nach einer halbstündigen Inkubation der Ansätze bei 11°C wurde die DNS wie beschrieben extrahiert, gefällt und in 6µl LoTE resuspendiert. Im Protokoll (Velculescu et al. 1997b) werden 37°C zur Inkubation empfohlen. Die Modifikation des Protokolls geschah aufgrund der Beobachtung, daß das Klenow Fragment bei einer niedrigeren Temperatur eine geringere 3'-5' Exonukleaseaktivität aufweist (mündliche Mitteilung A. Meisel) und somit eine höhere Wahrscheinlichkeit für den Erhalt längerer Tags besteht.

4.1.4.11 Ligation der Monotags zu Ditags

Um die Tags mit den Primern A und B zu amplifizieren, wurden die beiden Untergruppen A und B wieder zusammengeführt und End-zu-End ligiert, so daß Ditags einer Länge von ungefähr 100 Basenpaaren entstanden (Abb. 12). Von den mit Klenow begradigten DNS-Fragmenten wurden je 2µl der Lösungen (entspricht jeweils einem Drittel der Ansätze) der beiden Gruppen A und B zusammenpipettiert und die Ansätze auf 6µl mit 1,2µl 5x Ligase Puffer und 4 Units T4 Ligase ergänzt. Als Negativkontrolle für spätere PCR Schritte wurden die gleichen Ansätze ohne Ligase hergestellt. Die Ligation erfolgte bei 16°C über Nacht.

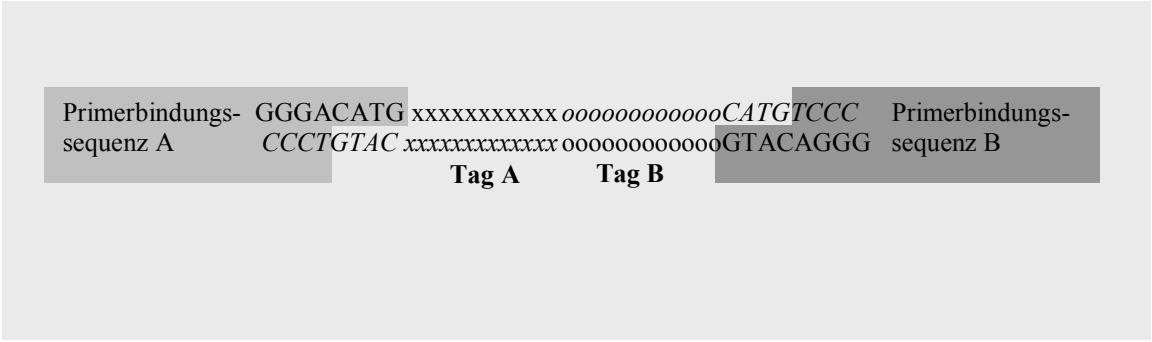

Abb. 12: Schema eines Ditags.

Um einen eventuell effizienteren Ansatz zu überprüfen (Lund et al. 1996), wurde jeweils ein zweiter identischer Ligationsansatz hergestellt und einem alternativen Ligationsprogramm, das zyklische Temperaturschwankungen aufwies, unterzogen: 99 mal 22°C/10" und 9°C/10", 99 mal 22°C/10" und 8°C/10", 12°C über Nacht. Das Ergebnis dieser beiden Ligationsansätze sollte nach der sich anschließenden PCR verglichen werden (siehe dort).

4.1.4.12 Amplifikation der Ditags mittels PCR

Um für die Sequenzierungen genügend Material zur Verfügung zu haben, müssen die Ditags amplifiziert werden. Der Vorteil einer PCR an dieser Stelle des SAGE-Durchlaufes liegt darin, daß ein durch diesen Amplifikationsschritt eingeführter quantitativer Bias dadurch kontrolliert werden kann, daß redundante Ditags, das heißt Ditags, welche wiederholt Tags einer bestimmten Kombination aufweisen, später von der Auswertung eliminiert werden können.

Den Ligationsansätzen wurde jeweils 14µl LoTE zugegeben und es wurden diverse Verdünnungen erstellt (1:10, 1:50, 1:100, 1:200). Die exakte Zusammensetzung der 50µl Standardansätze ist S. 40 (Methodenteil) zu entnehmen. Die PCR erfolgte im Warmstart-Verfahren und wurde mit folgendem Programm durchgeführt: zehn Sekunden 95°C; 24 bis 28 Zyklen: dreißig Sekunden 95°C, eine Minute 55°C und eine Minute70°C; abschließend fünf Minuten 70°C. Bevor die eigentlichen PCRs durchgeführt werden konnten, mußten die verschiedenen PCR Bedingungen jedoch optimiert werden. Hierbei wurde besonderes Augenmerk auf die Konzentration der Matrize, das heißt der Ditags, die Konzentration der dNTPs und die Anzahl der Zyklen gelegt.

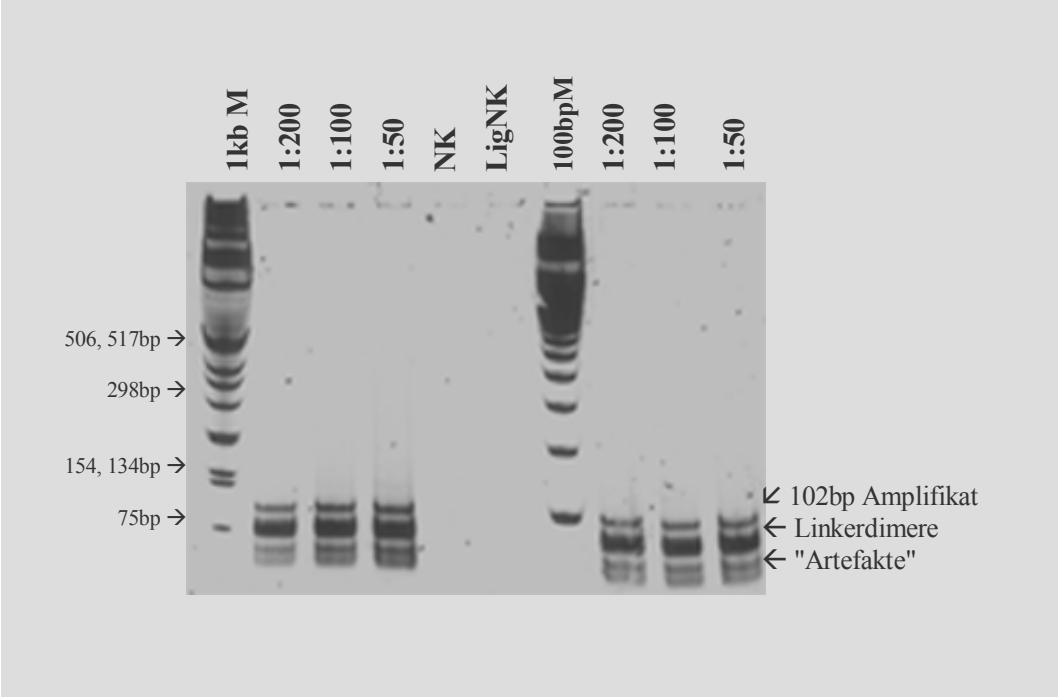

Abb. 13: K2 Ditag PCR in verschiedenen Verdünnungsstufen. 12% PAGE, ein Zehntel des PCR Ansatzes jeweils aufgetragen. NK: PCR Negativkontrolle (ohne Matrize), LigNK: Amplifikation der Negativkontrolle der Ligation (ohne Ligase). Selbst eine Verdünnung von 1:200 resultiert in die sichtbare Amplifikation der 102bp langen Ditags.

Optimierung der PCR 1

In Probedurchläufen von SAGE erwiesen sich eine Verdünnungsstufe von 1:400 (bei 28 Zyklen) und eine Zyklenzahl von 24 (Verdünnung von 1:50) noch als suffizient zur Amplifikation der 102bp Bande (keine Abbildung).

Optimierung der PCR 2

Hierbei wurden in beiden Kontrollgruppen sämtliche Verdünnungen der beiden Ligationsansätze getestet. Abbildung 13 zeigt, daß auch eine Verdünnung der Ligationsansätze von 1:200 (26 Zyklen) noch eine gut sichtbare 102bp Bande lieferte, und daß kein deutlicher Unterschied zwischen den beiden Ligationsprogrammen vorhanden war. Im Vergleich mit den Ergebnissen der Probedurchläufe (Optimierung 1, keine Abbildung) fiel auf, daß die Amplifikation des gewünschten 102bp Fragmentes effizienter als zuvor verlief. Als Erklärung dafür ist die Tatsache in Betracht zu ziehen, daß zum ersten Mal PAGE gereinigte Oligomere eingesetzt worden waren, während in den Probedurchläufen ungereinigte Primer verwendet worden waren.

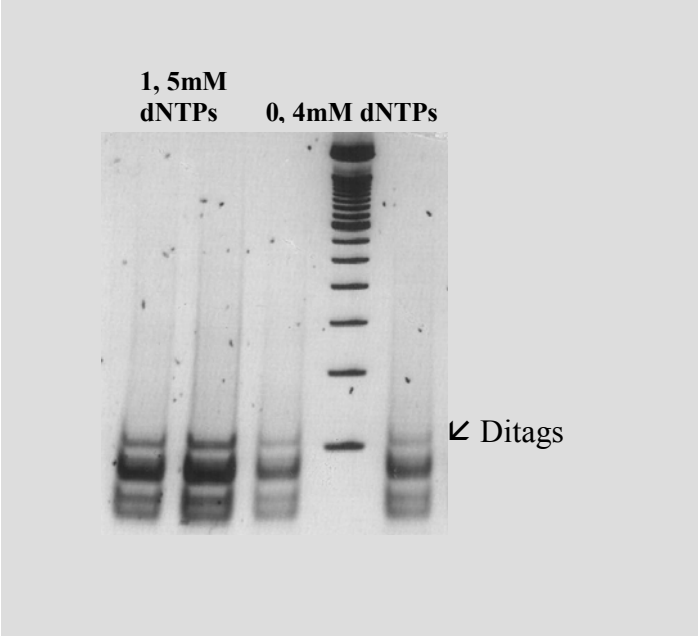

Abb. 14: K1 Ditag PCR. Vergleich zweier dNTP Konzentrationen. 8% PAGE. Die PCR mit unterschiedlichen dNTP Konzentrationen ergab für 0,4mM eine geringere Ausbeute als für 1,5mM. 100bp M: 100bp DNS Größenstandard.

Da auch die dNTP Konzentration das Resultat der PCR beeinflussen kann (siehe Diskussion), wurde diese probeweise ebenfalls variiert. Der Vergleich der standardmäßig verwendeten finalen dNTP Konzentration von 1,5mM mit einer Konzentration von 0,4mM (Abb. 14) ergab eine verringerte Ausbeute bei Verwendung der geringeren Konzentration, so daß die Standardkonzentration beibehalten wurde.

Die Negativkontrollen der PCR sowie der Ligation zeigten keine Amplifikation.

Um für die folgenden Schritte von SAGE genügend Material zur Verfügung zu haben, wurde letztendlich für alle weiteren PCRs eine Verdünnung von 1:100 und 26 Zyklen gewählt.

PCR Großansatz

Nach Abschluß der Optimierung der PCR Bedingungen wurden parallel je 50 50µl Ansätze durchgeführt.

Nach Gelreinigung (8% PAGE) der vereinigten Ansätze (Abb.15), Phenol-Chloroformextraktion und Ethanolfällung ließ sich anhand einer semiquantitativen DNS Mengenbestimmung mit Ethidiumbromid für K1 eine Gesamtmenge von 0,6µg und für K2 von 1,3µg schätzen. Da die zur Klonierung als notwendig erachtete Menge Ditag-DNS 10 - 20µg beträgt, wurden pro Kontrollgruppe 500 PCRs durchgeführt - statt wie im Protokoll angegeben 100, um diese Menge zu erhalten. Die PCR Großansätze erfolgten nach obigem Schema parallel. Nach erneuten Massenextraktionen von je 30 vereinigten Ansätzen ergab die semiquantitative Schätzung der Menge per Ethidiumbromid (vergleiche S. 34) eine ausreichende Gesamtmenge (K1: circa 25µg, K2: von 37µg).

4.1.4.13 Entfernung der Linker mittels NlaIII Verdau

Um die beiden Linkerduplexe wieder zu entfernen, wurde ein erneuter Verdau mit NlaIII durchgeführt. Durch diesen Schritt wurden kohäsive Enden produziert, was eine effiziente Ligation der Ditags miteinander und in einen Vektor ermöglichen sollte.

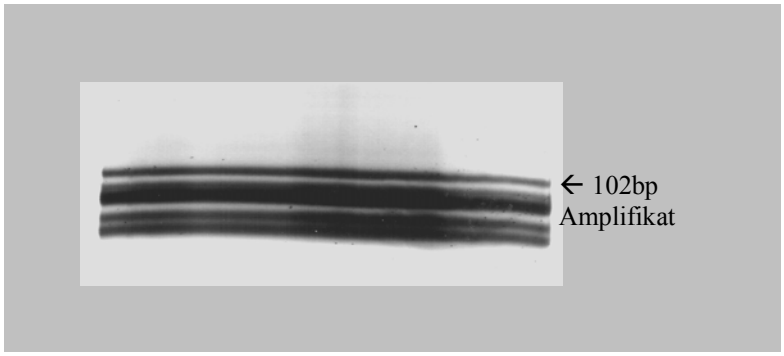

Abb. 15: Massenextraktion der 102bp Fragmente (K1). 8%PAGE, 30 PCR Ansätze.

In einem Volumen von 200µl wurden ungefähr 9µg (K1) beziehungsweise 10µg (K2) der gelgereinigten DNS und 150 (K1) beziehungsweise 170 (K2) Units NlaIII (circa 60ng DNS/U_{NlaIII}) angesetzt, so daß jeweils circa 30% des gesamten gelgereinigten Amplifikats eingesetzt wurden. Der Inkubation bei 37°C für drei Stunden schloß sich eine Phenol-Chloroform-Reinigung und eine Ethanolfällung an. Ein vorhergehender Verdau hatte ergeben, daß die von Velculescu et al. (1997b) angegebene Inkubationszeit von einer Stunde keine ausreichenden Ergebnisse (ohne Abb.) lieferte, so daß diese auf drei Stunden verlängert werden mußte. Eine semiquantitative Mengenabschätzung per Ethidiumbromid ergab nach

diesem Schritt 5,5µg DNS (K1) und 13,5µg (K2), gelöst in je 50µl LoTE. Die Gelreinigung der nun mehr 24 bis 26 Basenpaare langen Bande erfolgte nach elektrophoretischer Auftrennung im 12% PAGE (Abb. 16) wie beschrieben (S. 35 im Methodenteil). Um qualitativ hochwertige Dimere zu erhalten, wurde nur eine Spannung von 100 - 120V angelegt.

4.1.4.14 Ligation der Ditags zu Ketten

Um ein arbeits- und kostensparendes serielles Sequenzieren durchführen zu können, wurden die 26 bp langen Ditags zu Ketten aneinanderligiert, wobei die Erkennungssequenz von NlaIII (CATG) später als Interpunktion zwischen den einzelnen Ditags diente.

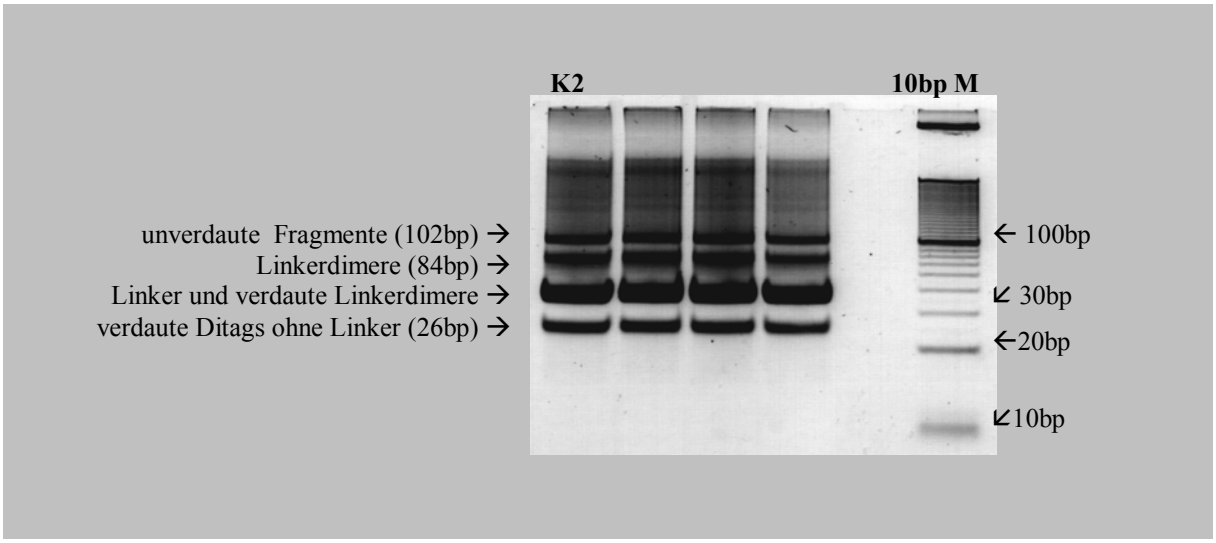

Abb. 16: Verdau der 102bp Fragmente mit NlaIII. 12%PAGE. Das gesamte Volumen (50µl≙ 180 PCRs) wurde auf vier Bahnen verteilt aufgetragen. 10bp M: 10bp DNS GS.

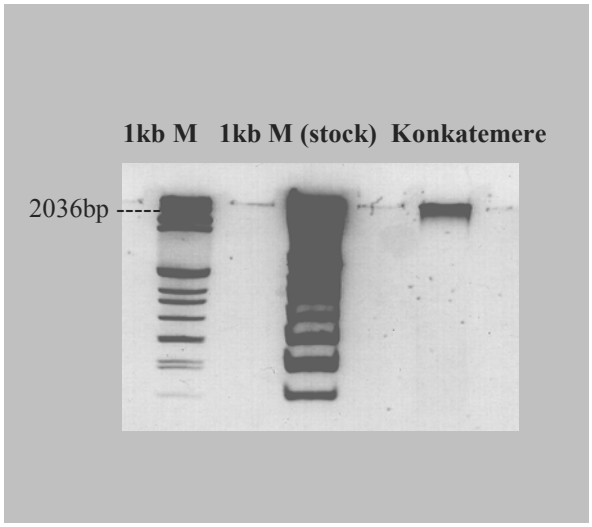

Abb. 17: K1 Konkatemerisierung (5U Li-gase, 1h). 12% PAGE (90V, 3h). Es wurde das gesamte Volumen (10µl) des Ligations-ansatzes aufgetragen. Es bildeten sich lediglich Ketten größer 2kb. 1kb M: DNS Größenstandard.

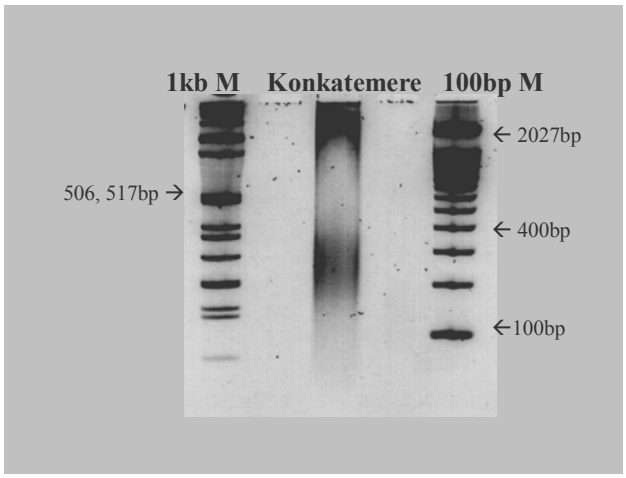

Abb. 18: Zweite K2 Ligation der Ditags zu Ketten (1U Ligase, 20min Inkubation). 8% PAGE, 10µl aufgetragen. 1kb/100bp M: DNS Größenstandards. Dieser Ligationsansatz ergab Ditagketten von circa 100 bis 4000bp.

Die gesamte Menge der gelextrahierten 26 bp Fragmente wurden per Zentrifugation konzentriert und in 7µl LoTE gelöst, was für K1 einem Drittel des per PCR gewonnenen Materials entsprach. Die Ditags wurden im 10µl Ansatz mit 5 Units T4 Ligase (GibcoBRL) versetzt. Es wurde bei 16°C für eine Stunde inkubiert. Velculescu et al. (1997b) empfehlen eine Inkubationszeit zwischen dreißig Minuten und zwei Stunden. Da sich in der anschließenden Gelelektrophorese (Abb. 17) zeigte, daß lediglich sehr lange Fragmente (größer 2000bp) entstanden waren, die für eine effiziente Ligation in einen Vektor wenig

geeignet gewesen wären, wurde das Protokoll dahingehend modifiziert, daß lediglich 1U T4 Ligase statt der vorgegebenen 5U eingesetzt wurde und die Reaktionszeit halbiert wurde. Für K2 (Verwendung von einem Sechstel des PCR Amplifikats) ergaben sich so nach der Gelelektrophorese Fragmentlängen von 450 - 4000bp (keine Abb.), welche in drei Größenstufen (450 - 600bp, 600 - 900bp und 900 - 4000) ausgeschnitten wurden und anhand der üblichen Vorgehensweise aufbereitet wurden. Um die Menge an klonierbaren Fragmenten noch weiter zu steigern, wurde in einem dritten Anlauf bei ansonsten gleicher Vorgehensweise die Ligationszeit weiter verkürzt (zwanzig Minuten), was Konkatemere (K2) im Bereich von 100 - 4000bp (Abb. 18) ergab. Auch hier wurden die Fragmente ausgeschnitten (400 - 800bp und 800 - 4000bp) und aufbereitet. Da weiterhin sehr lange Fragmente bei der Ligation der Ditags entstanden waren, wurde für die zweite Konkatemerisierung von K1 nach erneutem NlaIII Verdau von einem weiteren Drittel des PCR Amplifikats lediglich zehn Minuten lang bei 16°C inkubiert. Hiernach waren in der Elektrophorese Ditagketten im Bereich von 100 - 6000bp zu sehen (keine Abb.). Ausgeschnitten wurden drei Fraktionen, nämlich 500 - 700bp, 700 - 2000bp und 2000 - 6000bp. Die Aufbereitung erfolgte wie beschrieben (siehe S.48).

4.1.4.15 Klonieren der Ditagketten

Um die Tags sequenzieren zu können, wurden sie in einen Vektor (*pZErO*™ - 1, Invitrogen) ligiert und diese per Elektroporation in Bakterien übertragen. Verwendet wurden hierfür die ausgeschnittenen Konkatemere verschiedener Längenbereiche (400 - 4000bp).

Ligation der Ditags in einen Vektor

Als Vektor wurde ein SphI geschnittener *pZErO* verwendet. Nach dem Verdau von 1µg des Vektors (vergleiche S. 35) mit 2,5U SphI (NEB) wurde das Restriktionsenzym folgendermaßen inaktiviert: Inkubation bei 65°C für zwanzig Minuten, Zugabe von 90µl TE, Inkubation bei 70°C für sechs Minuten, dann bei Raumtemperatur für zehn Minuten. Die Endkonzentration des Vektors betrug 10ng/µl. Zur Ligation wurde das gesamte Volumen der gelgereinigten Konkatemere (5µl) im 10µl Ansatz mit 25ng des vorbereiteten Vektors und 2U der T4 Ligase (GibcoBRL) zusammengegeben und für vierzig Minuten bei 16°C inkubiert. Als Negativkontrolle wurde der Vektor alleine inkubiert, das heißt ohne Ditag- und Ligasezugabe. Dies diente auch der Überprüfung der Funktion des letalen Proteins, das bei Selbstligation des Vektors gebildet wird. Nach PC8 Extraktion und Ethanolfällung wurden die Ligationsprodukte in 8µl LoTE gelöst.

Elektroporation

Dies erfolgte wie dargestellt (siehe S. 37). Es wurden jeweils 1µl Ligationsansatz und 40µl kompetente *XL1 - Blue MRF E.coli* (Stratagen) benutzt. Ein Zehntel der transfizierten Bakterien wurde auf je einem Zeozin-haltigen Agarboden ausplattiert, so daß pro Ansatz zehn Platten zur Verfügung standen. Auf den Selbstligationsplatten so wie den Platten der Negativkontrolle (keine Ligase und keine Konkatemere im Ligationsansatz) ließ sich kein Wachstum beobachten.

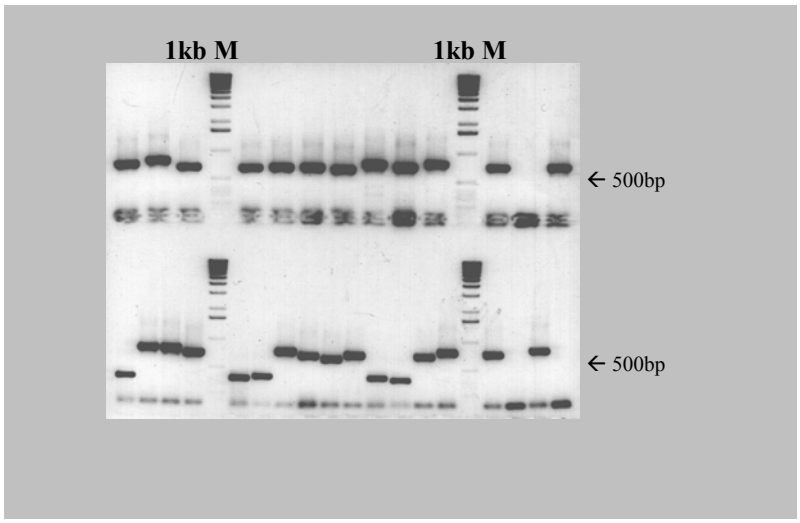

Abb. 19: Plasmidscreening PCR. 1% Agarosegel. Jeweils 8µl von 25µl aufgetragen. 1kb M: DNS Größenstandard.

Kontroll-PCR

Um zu überprüfen, ob die Länge der klonierten Tagketten den Erwartungen entsprach, wurden Screening-PCRs mit M13 F (universal) und R (reverse) Primern (Eurogentec) durchgeführt (Abb. 19). Für die PCR wurden 25µl Ansätze wie auf S. 49 beschrieben gewählt. In diesen Ansatz wurde eine mit einer sterilen Pipettenspitze gepickte Bakterienkolonie getaucht. Anschließend wurde nachfolgendes Programm durchgeführt: 95°C für zehn Minuten (initiale Denaturierung), 30 Zyklen mit 95°C für eine Minute, 60°C für eine Minute und 72°C für zwei Minuten, 72°C für zehn Minuten (finaler Elongationsschritt). Klone, welche eine Insertlänge größer als 500bp (inklusive 226 bp Vektorsequenz) aufwiesen, wurden anschließend zur Sequenzierung vorbereitet (siehe S. 47).

4.1.4.16 Sequenzieren

Die Erstellung von Pilotsequenzen erfolgte wie beschrieben (S. 48). Alle weiteren Sequenzierungen wurden in Kooperation mit dem Institut für Molekulare Biotechnologie in Jena nach dem dort verwendeten Standardprotokoll (siehe S. 49) durchgeführt.

4.1.4.17 Auswertung

4.1.4.17.1 Auswertung der Sequenzrohdaten

Die Auswertung der Rohdaten erfolgte am Institut für Molekulare Biotechnologie in Jena anhand von PHRED, einem Programm, das eine Qualitätsbewertung für jede Base vornimmt. Ein PHRED "quality score" von 30 beispielsweise bedeutet, daß die Base mit einer Genauigkeit von 99,9% gelesen wurde (International Human Genome Sequencing Consortium 2001). Es wurden zur Auswertung die im Rahmen des humanen Genomprojektes angewandten Qualitätskriterien benutzt.

4.1.4.17.2 Aufbereitung der sequenzierten Tagketten

Bevor die 10bp langen Tags einer Homologierecherche und der statistischen Auswertung unterzogen werden konnten, mußten sie aus den mehreren hundert Basenpaaren langen Sequenzen extrahiert werden. Dies geschah mittels der von K.W. Kinzler und Kollegen freundlicherweise zur Verfügung gestellten SAGE300 Software Version 3.01.

Bei der Analyse der verwertbaren Sequenzdaten per Computer konnten die Ditags anhand der Unterbrechungen durch die NlaIII Erkennungssequenz (CATG) identifiziert werden. Da aufgrund der End-zu-End Ligation der Monotags das zweite Tag sozusagen rückwärts in 3'→ 5' Richtung sequenziert worden war, wurde dieses jeweils von der Software umgedreht, nachdem die Grenze der beiden Tags voneinander anhand anzugebener Tag- und Ditaglängen bestimmt worden war.

Vor der automatisierten Analyse mußten unter "Präferenzen" grundlegende Einstellungen angegeben werden. Aus einer *'drop down'* Liste von 92 Typ II Restriktionsendonukleasen konnte das verwendete Verankerungsenzym ausgewählt werden. Des weiteren mußte die geschätzte Taglänge (abhängig von den verwendeten Enzymen) mit 4 bis 13bp angegeben werden - wobei empfohlen wird, ein Basenpaar weniger als erwartet zu wählen (also 10bp), um bei schwankenden Ditaglängen bei der Zuordnung der Basen zu einem Tag auf der sicheren Seite zu sein und auch kürzere Tags in die Auswertung miteinbeziehen zu können. Die erforderliche Angabe der maximalen Länge der Ditags ergab sich aus der zweifachen

Taglänge, zu welcher mindestens 2bp addiert werden, um auch längere Ditags in die Analyse einschließen zu können. Als Mindestlänge wurde von dem Programm automatisch die zweifache Monotaglänge eingesetzt. Außerdem war erforderlich, den Organismus, aus welchem das analysierte Gewebe entnommen worden war, einzutragen (hier: *Mus musculus*). Eine Verknüpfung stellte die Verbindung zu einem sogenannten '*Tag Exclude Manager*' her. Dieser beinhaltet eine Liste von maximal vierzig Tags, welche von der Analyse ausgeschlossen werden sollten, wie zum Beispiel die Linkersequenzen (vergleiche Tabelle 6).

```
Date Analyzed: 06-22-1999
Project Name: SAGE10K1
Enzyme: NlaIII-CATG        Taglength:  10
Maximum DiTag length: 24
Project Files:  650        Total Tags:  9876
Project Duplicate Dimers:  169

Input File: D:\SAGE MAUS\SAGE10K1\CLKIaa14d11.t1.seq
Start position:  1         Stop position:  506
Sequence Length:  0

            1 )        ACTGCTTTTCCCTTGGTAAGTAA
            2 )        GATTAAAAGCCTTAATAGGGA
            3 )        TCCCCGTACACCTCCAGGTCAGA
            4 )        TGGACACTCAAAGCCGGCGGCG
            5 )        TCCCTATTAAGGGCAAGTTGTGC
            6 )        AAACCCTCTNTTGACAAGTTTC
            7 )        TTCCGTGAACCGTACGTAACCA
            8 )        TCCCTATTAAGCCCGGGGTCA
            9 )        Dimer too long.  Length =  141

Total Dimers:  9
Short Dimers:  0
Long Dimers:  1
Duplicate Dimers:  0
Good Tags:  15

Project Files:  650
Project Tags:  9876
Project Duplicate Dimers:  169
```

Abb. 20: Ausgabe der Analyse eines Konkatemers mit SAGE300.

Linkerderivate	Fehlerart	Linkerderivate	Fehlerart	Linkerderivate	Fehlerart
TCCCTATTAA-GCC	Orginalsequenz	TCCATATTAA	sub	TCCCTACTAA	sub
ACCCTATTAA	sub	TCCGTATTAA	sub	TCCCTAGTAA	sub
CCCCTATTAA	sub	TCCTTATTAA	sub	TCCCTATAAG	del
GCCCTATTAA	sub	TCCCAATTAA	sub	TCCCTATAAA	sub
CCCTATTAAG	del	TCCCCATTAA	sub	TCCCTATCAA	sub
TACTATTAAG	del+sub	TCCCGATTAA	sub	TCCCTATGAA	sub
TGCCTATTAA	sub	TCCCATTAAG	del	TCCCTATAAG	del
TTCCTATTAA	sub	TCCCTCTTAA	sub	TCCCTATTCA	sub
TCCTATTAAG	del	TCCCTGTTAA	sub	TCCCTATTGA	sub
TCACTATTAA	sub	TCCCTTTTAA	sub	TCCCTATTTA	sub
TCGCTATTAA	sub	TCCCTTTAAG	del	TCCCTATTAG	del?sub?
TCTCTATTAA	sub	TCCCTAATAA	sub	TCCCTATTAC	sub

Tabelle 6. Liste der Linkerartefakte (Linker 1B), die von der SAGE300 Software als Vorlage zur Elimination von Linkersequenzen verwendet wurde. Die Liste für Linker 2B verhält sich analog. Bei den von der Orginalsequenz (um eine Base) abweichenden Sequenzen gibt die Spalte "Fehlerart" eben diesen an: "sub" für Substitutionen und "del" für Deletionen.

Nach der Eingabe dieser Informationen konnte die Analyse des SAGE Projektes vorgenommen werden. Hierzu wurde ein Ordner erstellt, welcher sämtliche Informationen zum aktuellen Projekt enthält. Die Sequenzen wurden im selben Verzeichnis mit der Extension "*.seq" abgespeichert. Um sicher zu gehen, daß es sich um relevante Daten handelt, wurden die Sequenzen ansatzweise vor der automatisierten Auswertung "per Hand" gegengelesen. Pro Klon war die Auswertung von 9000bp möglich. Des weiteren wurde die Position der Start (1 - 1000)- und Stoppbase (<9000) angegeben (in den vorliegenden Daten meist knapp 500bp). Auf diese Weise können zum Beispiel Strecken, welche Vektorsequenzen aufweisen, von der Analyse ausgeschlossen werden. Durch die automatische Auswertung wurden die einzelnen Tags extrahiert und das gesamte Projekt auf mehrfach auftretende Ditags und Linkersequenzen (siehe Tabelle 6) überprüft, wobei auch Ditags in umgekehrter Orientierung aussortiert wurden. Ditags, welche zu kurz waren, sowie Ditags, welche als zu lang erkannt wurden, wurden von der Auswertung ebenfalls ausgeschlossen (siehe Abb. 20). Die SAGE300 Software bietet auch die Möglichkeit der statistischen Auswertung per Monte Carlo Simulation (ohne Teststatistik), so wie des Vergleichs verschiedener Projekte. Auf diese Option soll im Rahmen der Diskussion der statistischen Probleme von SAGE eingegangen werden.

4.1.4.17.3 Sekundäre Elimination verbliebener Linker

Bei der Durchsicht der beiden durch den vorherigen Schritt entstandenen Taglisten fiel auf, daß sich darunter Sequenzen befanden, welche den eingesetzten Linkern stark ähnelten, jedoch nicht von der SAGE300 Software aussortiert worden waren. Um diese Sequenzen auch noch zu eliminieren, wurde in Kooperation mit Dr. Oliver Redner[5] ein Programm in C/C++ erstellt. Dieses ist in der Lage, Tags zu erkennen, welche sich um eine Base von häufigeren Tags unterscheiden. Damit orientiert es sich an der Vorgehensweise der SAGE300 Software. Die Taglisten wurden von Excel in ".prn" Textdateien konvertiert und insofern editiert, als die beiden originalen Linkersequenzen als 10er Tag mit dem fiktiven Häufigkeitswert von 700 an den Anfang gesetzt wurden. Nach dem Einlesen der Tagliste wurde der maximale Wert für Tags, die als Matrizen dienen sollten, nämlich die beiden Linkersequenzen, auf 400 festgelegt. Auf diese Weise waren die beiden Linkertags die Grundlage für den Abgleich mit allen Tags des Projekts (maximaler Häufigkeitswert 379). Das Ergebnis ist **Tabelle 7** zu entnehmen. Es wurden in K1 31 zusätzliche Linkerartefakte, die zusammen 364 mal vorkamen, auf diese Weise identifiziert und eliminiert. In K2 waren dies 24 verschiedene Tags (Masse: 269).

[5] Institut für Mathematik und Informatik, Universität Greifswald, Friedrich-Ludwig-Jahn-Str. 15A, 17487 Greifswald.

Linkerderivate	K1	K2	Linkerderivate	K1	K2
TCCCCCGTAC	75	60	TCCCCGACAC	3	1
TCCCCTATTA	59	37	TGCCCTATTA	2	0
TCCCTTATTA	26	24	TCCCATATTA	2	3
TCCCGTACAC	22	18	TACCCCGTAC	2	0
TCCCCGTAAC	22	12	CCCCGTACAC	2	3
TCCCCGTACC	20	18	TCCTCTATTA	1	0
TCCCCGTTAC	20	11	TCCTCCGTAC	1	0
TTCCCTATTA	20	9	TCCCTTTAAA	1	0
TCCCTAATTA	19	11	TCCCTACTTA	1	0
TCCCCGGTAC	17	13	TCCCGTATTA	1	1
TTCCCCGTAC	12	16	TCCCCGTCAG	1	1
ATCCCTATTA	9	3	TCCCCGTAAG	1	0
TCCCCGTCAC	7	15	TCCCCGACAG	1	0
ATCCCCGTAC	7	4	CCCTATTAAA	1	0
TCCCGTACAG	3	0	TCTCCCGTAC	0	1
TCCCCTACAC	3	3	TCCTATTAAA	0	1
TCCCCGTACG	3	3	TCCCCGACAA	0	1

Tabelle 7. Sekundär eliminierte Linkerartefakte. In den Spalten ist nach den beiden Gruppen getrennt die jeweilige Häufigkeit angegeben.

4.1.4.17.4 Sequenzfehlerkorrektur

Um eventuelle Sequenzfehler der in der vorliegenden Arbeit erstellten Expressionsprofilen näherungsweise, aber systematisch (ausführliche Erklärung siehe S. 130) - zu korrigieren, wurde erneut das oben beschriebene Programm in C/C++ eingesetzt. Diesmal wurden alle Singletons, die sich um eine Base von häufiger auftretenden Tags unterschieden und damit vermutlich Resultat eines Sequenzierfehlers waren, diesen zugeordnet. Dies führte dazu, daß die Menge der häufiger auftretenden Tags sich um ein Tag erhöhte, während das als fehlerhaft interpretierte Singleton aus dem Datensatz verschwand. Als Grenze für die Tags, welchen die Singletons zugeordnet werden sollten, wurde das Kriterium ">= 3" gewählt. In dem Fall, daß ein zuzuordnendes Tag sich von mehreren öfter auftretenden Tags um eine Base unterschied, wurde es dem häufigsten zugewiesen. Falls es dabei mehr als ein Tag mit einer bestimmten Häufigkeit gab, wurde es dem in der Liste oben stehenden zugeordnet. Dies bedeutet, daß hier Kriterien der Listensortierung (Mittelwert, Alphabet) zur Anwendung kamen, die inhaltlich nicht mit der Sequenzierung und ihren Fehlern zusammenhängen. Doch ist dies in

Ermanglung besserer Vorgehensweisen nicht zu umgehen und traf lediglich in 1,0% (K1) beziehungsweise 1,8% (K2) der Fälle zu. Wenn Taghäufigkeiten sehr nah beieinander liegen (beispielsweise 3 und 4) ergibt sich das gleiche Problem. Dies wurde jedoch nicht berücksichtigt, da es den Rahmen der Analyse gesprengt hätte. Auf diese Weise wurden in K1 1140 Singletons zugeordnet und in K2 1163. In beiden Gruppen entspricht dies 8,4% der Gesamttagmenge.

4.1.4.17.5 Homologierecherche

Die Zuordnung der extrahierten Tags zu bereits bekannten Gensequenzen erfolgte unter Verwendung einer über das Internet erhältlichen Gendatenbank, welche mit Hilfe des UniGene Projektes (http://www.ncbi.nlm.nih.gov/UniGene) speziell für SAGE erstellt worden war (ftp://ncbi.nlm.nih.gov/pub/sage) (Lash et al. 2000). UniGene bündelt ähnliche GenBank Sequenzen in Gruppen (Cluster), welche jeweils ein Gen repräsentieren. Um dieses Projekt für SAGE nutzbar zu machen, waren diese Cluster folgendermaßen bearbeitet worden: Sortierung nach Spezies, Orientierung der Sequenzen anhand eines Polyadenylierungssignals (ATTAAA/AATAAA) oder -schwanzes (mindestens acht A) oder einer vorhandenen Angabe zur Orientierung, Extraktion eines SAGE entsprechendem 10bp Tags unmittelbar 3' der am meisten 3' liegenden NlaIII Erkennungssequenz, Zuordnung der UniGene Cluster Nummer zu dem Tag. Nicht jedes potentielle Tag konnte mit einem Cluster gepaart werden, der aus gut charakterisierten cDNS Sequenzen besteht, deren Sequenzierfehler vernachlässigbar ist. Häufig fanden sich in der Datenbank lediglich ESTs, welche nur einmal sequenziert worden waren und somit einen geschätzten Fehler von 10% für die 10bp lange Sequenz (Lash et al. 2000) aufweisen. Das bedeutet, daß 10% der Tag-UniGene Cluster Paarungen aufgrund dieses Fehlers entstanden waren. Um diesem Problem zu begegnen, wurden 10% der seltensten Tag-UniGene Zuordnungen entfernt, da hier Fehler am wahrscheinlichsten sind. Das Ergebnis dieser Prozesse ist oben erwähnte Datenbank, welche sämtliche als zuverlässig bezeichnete UniGene Cluster - bestehend aus cDNS Sequenzen und ESTs - enthält. Die dort gespeicherten ESTs werden in vier Klassen eingeteilt, wobei ihr Wert für die Tag-Gen-Paarungen von oben nach unten abnimmt. Als erstens kommen ESTs, die als 3' orientiert in der GenBank geführt werden und ein Polyadenylierungssignal oder -schwanz aufweisen können. Es folgen ESTs ohne Orientierungsangabe, aber mit Poly(A)-Kennzeichen, ESTs mit einer 5'Orientierung und Poly(A)-Kennzeichnung und als letztes ESTs mit 3'Orientierung ohne Hinweis auf eine Polyadenylierung (siehe Tabell 8 Spalte "Ursprung"). Diejenige Datei, welche *M.musculus* Sequenzen enthielt und NlaIII als Verankerungsenzym verwendet hatte ("SAGEmap_tag_ug-

rel-Nla3-Mm"), wurde von dem NCBI FTP Server heruntergeladen und als MS Access Tabelle gespeichert. Per Auswahl-Abfrage erfolgte eine Verknüpfung mit den erstellten Tagsequenzen. Alternativ wurden Zuweisungen anhand des '*Tag Mapping*' Werkzeugs der SAGE Internetseite (http://www.ncbi.nlm.nih.gov/sage) vorgenommen, das auf demselben Prinzip wie die Datenbank basiert. Das Ergebnis dieser Recherche ist für die fünfzig häufigsten Gene des hier vorliegenden Projektes zu entnehmen.

UniGene Cluster Nr.	**Tagsequenz**	**11. Base**	**Beschreibung**	**Ursprung**	**Mittelwert (Tags)**
keine reliable Zuordnung vorhanden	GCTGCCCTCC	A	mitochondriale Sequenz	EST: 3'Orientierung, Poly(A)-Kennzeichnung	332,5
35 reliable Clusterzu-ordnungen vorhanden	GTGGCTCACA	A	Beispiel: Mm.100791: RIKEN cDNS 2700038G22 Gen	cDNS	157
keine zuverlässige Zuordnung vorhanden	ATACTGACAT	T	mitochondriale Sequenz	EST: 3'Orientierung, Poly(A)-Kennzeichnung	106
keine zuverlässige Zuordnung vorhanden	AGGAGGACTT	A	mitochondriale Sequenz	EST: 5'Orientierung, Poly(A)-Kennzeichnung	92,5
keine zuverlässige Zuordnung vorhanden	AACGGCTAAA	C	mitochondriale Sequenz	EST: 3'Orientierung, Poly(A)-Kennzeichnung	76,5
	TCCCCCGTAC	A	Linkerartefakt		67,5
keine zuverlässige Zuordnung vorhanden	AGGACAAATA	T	mitochondriale Sequenz	EST: keine Orientierungsangabe, Poly(A)-Kennzeichnung	65
keine zuverlässige Zuordnung vorhanden	ATGACTGATA	A	mitochondriale Sequenz	EST: 3'Orientierung, Poly(A)-Kennzeichnung	64
keine zuverlässige Zuordnung vorhanden	ATAATACATA	A	Mm.14087: Komplement Komponente 4 bindendes Protein	EST: 3'Orientierung, Poly(A)-Kennzeichnung	61,5
27 zuverlässige Zuordnungen gefunden	AAAAAAAAAA	A	1. Mm.104540: RIKEN cDNS 1500039N14 Gen	cDNS	57,5
keine zuverlässige Zuordnung	AGCAATTCAA	A	mitochondriale Sequenz	EST: 5'Orientierung, Poly(A)-Kennzeichnung	50

vorhanden					
Mm.2992	GCTTCGTCCA	G	Myelin Basic Protein	cDNS	48,5
Mm. 30245	TCCCCTATTA	A	Phosphatidylserine Decarboxylase Klon MGC:7133 ähnlich	cDNS	48
keine zuverlässige Zuordnung vorhanden	AGCAGTCCCC	T	mitochondriale Sequenz	EST: 3'Orientierung, Poly(A)-Kennzeichnung	44
	CAAACCTCCA	T	keine Homologie gegenwärtig		38,5
Mm. 5289	GCCTCCAAGG	A	Glyceraldehyde-3-Phosphat Dehydrogenase	cDNS	35,5
Mm. 4263	CCTTGCTCAA	T	Cystatin C	cDNS	30,5
Mm. 29846	CCGCCCCTTT	C	verwandt einer "N-myc downstream 1" regulierten Sequenz	cDNS	30
	TCCCTTATTA	A	Linkerartefakt		25
Mm. 196614	AGGCAGACAG	T	Eukaryotischer Translationselongations- Faktor 1 alpha 1	cDNS	24
Mm. 4881	GCGGGGTCGC	C	Granin-ähnlicher neuroendokrinischer Peptidvorläufer	cDNS	23
Mm. 44101	GCCCCCCTCT	C	schwach ähnlich der I49143 gastrischen H(+)-K(+)-ATPase Alpha Subunit (Mm)	EST: 5'Orientierung, Poly(A)-Kennzeichnung	22,5
Mm. 29807	GCGCCAGCTC	A	Ubiquitin Carboxy-terminale Hydrolase L1	cDNS	22
Mm. 18041	GCACAACTTG	C	Calmodulin II	cDNS	22
Mm. 39185 Mm. 21110	GCTGCCCCTC	C	Mm. 39185: EST Mm. 21110: EST	EST: 3'Orientierung, Poly(A)-Kennzeichnung bzw. ohne Orientierung, Poly(A)-Kennzeichnung	22
Mm. 1268	AAATTATTGG	G	Proteolipid Protein (Myelin)	cDNS	21,5
Mm. 5246	GAGCGTTTTG	G	Peptidylprolylisomerase A	cDNS	20,5
	TCCCGTACAC	G	Linkerartefakt		20
keine zuverlässige Zuordnung vorhanden	ACCAATGAAC	A	1. Mm. 150211: schwach ähnlich dem T21052 hypothetischen Protein F226125 (C.elegans) 2. Mm. 22575: Melanomantigen, Familie D, 2 3. Mm. 4962: differentiell exprimierter Tumor	EST: 3'Orientierung, Poly(A)-Kennzeichnung	20
Mm. 34246	ACAAACTTAG	G	Calmodulin	cDNS	19,5
Mm. 4024	GAAGCAGGAC	C	nicht muskuläres Kofilin 1	cDNS	19
	TCCCCGTACC	A	Linkerartefakt		19
9 zuverlässige Zuordnung gefunden	CCTTTAATCC	C	Beispiel: Mm. 1007: Proteasom (Prosom, Makropain) Subunit alpha Typ 3	cDNS	19
Mm. 30155	CGTCTGTGGA	G	lysosomale ATPase, H+ transportierend (vakuoläre Proton Pumpe) 16kD	cDNS	19

Mm. 13859	AGAGCGAAGT	G	RIKEN cDNS 1810055P16 Gen	cDNS	18,5
Mm. 13020	AGGTCGGGTG	G	Ribosomales Protein L13a	cDNS	18,5
Mm. 196396	GCTGCCCTAG	A	Alpha 1 Tubulin (M α 6)	cDNS	18
Mm. 297	CCCTGAGTCC	A	Melanom X-Actin	cDNS	17,5
	TCCCCGTAAC	A	Linkerartefakt		17
Mm. 1240	AAGTGTCGCC	G	Wachstumshormon	cDNS	16,5
Mm. 29857	CAGCTCTGCC	T	Neurogranin (Protein- Kinase C Substrat, RC3)	cDNS	16
Mm. 43005	TGACCCCGGG	A	Fusionsprodukt 1 des Ubiquitin A-52 Rest ribosomalen Proteins	cDNS	16
	TCCCCGTTAC	A	Linkerartefakt		15,5
Mm. 42829	TTTCCAGGTG	T	muskuläres Selenoprotein W	cDNS	15,5
Mm. 19605 Mm. 5290	GCCCGGGAAT	A	1. Hexokinase 1 2. ribosomales Protein L17	cDNS	15,5
Mm. 1008	GTGACCTGGC	C	zerebrale Prostaglandin D2 Synthase (21 kDa)	cDNS	15,5
	TCCCTAATTA	A	Linkerartefakt		15
	TCCCCGGTAC	A	Linkerartefakt		15
Mm. 3158	GGCTTCGGTC	T	ribosomales Protein P1	cDNS	15
Mm. 1147	ATCCGCACCC	T	murines Calmodulin III (3'UTR)	cDNS	15

Tabelle 8: Resultate der Homologiesuche für die 50 häufigsten Gene. Falls einer Tagsequenz mehr als 2 Gene zugeordnet werden konnten, wurde nur eines beispielhaft in die Tabelle aufgenommen. Die Spalte "Ursprung" verweist auf die Herkunft der Gensequenzen, wobei cDNS' am besten charakterisiert sind und ESTs in absteigender Rangordnung folgen: 1. 3'Orientierung mit Poly(A)-Kennzeichnung, 2. keine Orientierungsangabe, aber Poly(A)-Kennzeichnung, 3. 5'Orientierung mit Poly(A)-Kennzeichnung, 4. 3'Orientierung ohne Poly(A)-Kennzeichnung. In der Rubrik "Mittelwert" befinden sich die arithmetischen Mittel der Tags von K1 und K2. Da die Homologierecherche vor der sekundären Elimination verbliebener Linkerartefakt durchgeführt worden war, sind ebensolche enthalten.

4.1.4.18 Quantitatives Resultat

Insgesamt wurden nach Abzug von 890 replikativen Ditags (entspricht 5,33% von 33406 Tags) 31626 Tags lesbar und auswertbar sequenziert. Es befanden sich darunter nach Analyse mit der SAGE300 Software 3494 Sequenzen (11,05%), welche Linkern entsprachen und somit nicht gewertet werden konnte. Die sekundäre Analyse der Daten mit dem bereits erwähnten Programm zur Elimination von Linkerartefakten (Originalsequenzen und Sequenzen, die sich um eine Base unterscheiden) ergab zusätzlich in K1 364 Linkerartefakte und in K2 269. Insgesamt wurden folglich 14,39% Linkerartefakte gefunden. Somit konnten 13584 (K1) und 13915 (K2) Tags ausgewertet werden.

Es wurden insgesamt 14159 verschiedene Transkripte detektiert. Sie verteilten sich folgendermaßen auf die beiden Gruppen: K1 8302 und K2 8054 unterschiedliche Transkripte. Nur 269 dieser Tags traten öfters als oder genau zehnmal in beiden Gruppen zusammen auf

(Mittelwert: 5). Tags, welche nur einmal gezählt wurden (Singletons), gab es insgesamt betrachtet 12846, wobei K1 6570 Singletons aufwies, was bei 8302 verschiedenen Transkripten 79,1% entsprach, und K2 6276 (75,6% von 8054 verschiedenen Transkripten). Das bedeutet, daß jeweils gut $^3/_4$ der detektierten Gene in einem Expressionsbereich von 20 Kopien pro Zelle lag. Dabei wurde der Umrechnung der gezählten Tags in die Menge der in einer Zelle enthaltenen Transkripte eine Gesamtzahl von 300000 Transkripten pro Zelle zugrunde gelegt (Hastie und Bishop 1976).[6] Fast 99% der detektierten Gene hatten ein Expressionsniveau von weniger als 200 Kopien pro Zelle (vergleiche **Tabelle 9** Spalte "Gene" und Abb. 21).

Kopien pro Zelle	K 1 Gene	Menge	K 2 Gene	Menge
> 2000	0,04%	3,95%	0,05%	5,39%
2000 - 200	0,99%	11,97%	1,25%	14,11%
200 - 20	98,98%	84,08%	98,70%	80,50%
< 20	nicht beurteilbar			

Tabelle 9. Verteilung der Häufigkeiten in Klassen (in Prozent). Unter der Rubrik "Gene" ist der Anteil der verschiedene Gene, welche in der jeweiligen Häufigkeitsklasse erscheinen, in Prozent aufgeführt, unter "Menge" der zugehörige Anteil, den diese Gene an der Gesamtmasse der gezählten Kopien haben. Bei einer Anzahl von pro Gruppe ungefähr 14000 sequenzierten Transkripten liegen Gene, welche weniger als 20 Kopien/Zelle aufweisen, außerhalb des Meßbereichs. Um beispielsweise Transkripte mit 1 Kopie/Zelle wiedergeben zu können, müßten 300000 Tags sequenziert werden.

Die Bandbreite der detektierten Transkripthäufigkeiten reichte bis zu 8200 Transkriptkopien pro Zelle. Bei Betrachtung von deren Verteilung (**Tabelle 9** Spalte "Menge"), zeigt sich, daß die geringste Expressionsklasse die Masse (80%) der gefundenen Transkripte ausmacht. Dieser Befund läßt sich auch dem '*Scatterblot*' entnehmen (siehe Abb. 21). Im zweiten Abschnitt der Arbeit (ab S. 112) soll diese Graphik durch statistische Aufarbeitung verfeinert werden.

[6] x Kopien pro Zelle / 300000 = Anzahl der spezifischen Tags / Tagsgesamtzahl der Gruppe.

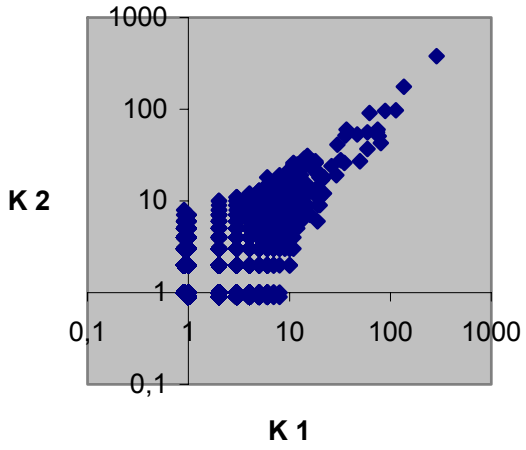

Abb. 21: Scatterblot von K1 gegen K2. Auf beiden Achsen ist logarithmisch die Anzahl der gezählten Tags aufgetragen. Genen, welche in einer der beiden Gruppen nicht auftraten, wurde zur Darstellung der Wert 0,9 zugewiesen. Die Abbildung macht deutlich, daß die Masse der Gene geringe Tagzahlen aufweist.

4.2 Diskussion der Etablierung von SAGE

Die vorliegende Arbeit ist im Rahmen eines Projektes zur Untersuchung der Genexpression bei Tiermodellen neurologischer Erkrankungen entstanden. Mit herkömmlichen Kandidatenansätzen und den entsprechenden Methoden wie beispielsweise Northern Blotting ist eine Expressionsanalyse jedoch nur in beschränktem Umfang zu realisieren. Ziel war daher die Etablierung eines Verfahrens wie SAGE, das die Analyse des gesamten zerebralen Transkriptoms zuläßt. Die eingangs gestellte Frage, ob eine derart komplexe Methode, die bisher weltweit nur in einer begrenzten Anzahl von Laboren erfolgreich durchgeführt werden konnte (Yamamoto 2001), in einem molekularbiologischen Standardlabor etabliert werden kann, kann anhand der exemplarischen Durchführung von SAGE an gesunden Mäusegroßhirnen positiv beantwortet werden.
Im folgenden sollen methodische Probleme und Eigenheiten der Durchführung und Auswertung von SAGE sowie die durchgeführten Modifikationen der Originalmethode diskutiert werden.

4.2.1 Methodische Probleme der Durchführung von SAGE

4.2.1.1 Kontamination mit Linkersequenzen

Die Rate an Linkersequenzen, die sich nach Auswertung mit der SAGE300 Software unter den sequenzierten Tags befinden, lag in den vorliegenden Daten mit 11% deutlich höher als in der Literatur angegeben. Mit dem Ausschluß aller möglicher Linkerartefakte[7] mittels des extra entwickelten Programms (siehe S. 73) stieg dieser Anteil in der Kontrollgruppe 1 auf 14,4% und in der Kontrollgruppe 2 auf 11,7% an. Velculescu et al. (1997) sprechen von einer Kontaminationsrate von 3,7% (bei 62965 Tags), Kal et al. (1999) geben eine Rate von 5,27% (bei 10943 Tags) und 6,08% (bei 3847 Tags) an und Welle et al. (1999) eine von 1,35% (bei 53875 Tags). Allerdings wird die Rate der Linker nicht in allen Publikationen angegeben, so daß in der oben zitierten Literatur ein gewisser Bias zugunsten geringer Kontaminationsraten vorliegen könnte. Zudem stellt sich die Frage nach den dort angewandten Eliminationskriterien (die in keiner Publikation genannt werden) und dem damit vorhandenen Einfluß auf die Zahl der als Linkerartefakt deklarierten und aussortierten Tags.
Ein hoher Anteil an Linkersequenzen in den Polytagketten wie in den vorliegenden

[7] Dies beinhaltet die originalen Linkersequenzen sowie Sequenzen, die sich um eine Base von diesen unterscheiden.

Ergebnissen ist deswegen nachteilig, weil dadurch die Sensitivität von SAGE als Meßmethode absinkt. Wenn die Kosten eines Projekts, das heißt die sequenzierte Menge, konstant gehalten werden sollen, verringert sich durch die Kontamination mit Linkerartefakten die Anzahl der auswertbaren Tags und somit die Sensitivität. Zudem können inkorporierte angeschnittene Linker(teile) die Effizienz eines SAGE-Durchlaufes senken, indem sie zum Beispiel die Ligation der Ditags zu Polytags stören, da sie zu Kettenabbrüchen führen können (siehe S. 86).
Angesichts dieser relativ hohen Anzahl an unspezifisch inkorporierten Linkersequenzen wäre für einen erneuten Durchlauf von SAGE zu empfehlen, den Waschschritt, welcher sich an die Ligation der Linker an die immobilisierten cDNS Fragmente anschließt, durch mehrmalige Durchführung zu optimieren, um sämtliche ungebundenen und damit überzähligen Linker zu entfernen und auf diese Weise von der Weiterverarbeitung auszuschließen.
Ein weiterer Ansatz zur Lösung dieses Problems findet sich bei Angelastro et al. (2000a). Dort wird nur ein Fünftel der im Originalprotokoll angegebenen Menge an Linkern eingesetzt. Dies hat zur Folge, daß die Menge der nicht an cDNS Fragmente ligierten Linker geringer wird und damit die Kontaminationsrate sinkt.

4.2.1.2 Amplifikation der Ditags

Um die zu Weiterverarbeitung erforderliche Menge zu erhalten, wurden - anders als im Protokoll angegeben (100 parallele PCR Ansätze) - pro Kontrollgruppe 500 PCRs durchgeführt. Die Notwendigkeit, diese Modifikation vorzunehmen, mag darin ihre Ursache haben, daß Amplifikation der Ditags nicht optimal eingestellt werden konnte. Dies ist daran zu sehen, daß die 80bp Fragmente (Linkerdimere) wesentlich effizienter amplifiziert wurden als die Linkerditagkomplexe (Abb. 13), während Velculescu et al. (1997b) fordern, daß diese Bande von höchstens gleicher Intensität wie die darüber laufende 102bp Bande sein soll.
Die Problematik dieses SAGE Schrittes liegt darin, eine maximale Ausbeute an spezifischen PCR Produkten bei möglichst großer Diversität der Tags zu erreichen. Kritische Parameter sind dabei die Konzentrationen der eingesetzten DNS Matrize und der Nukleotide sowie die Anzahl der Amplifikationsschritte.
Bei einer zu geringen Menge an Matrize ist die Ausbeute nicht maximal. Zu viel DNS pro Einzelansatz reduziert ebenfalls die maximal mögliche Endmenge, da insgesamt nur eine begrenzte Menge an cDNS, die als Matrize eingesetzt werden kann, zur Verfügung steht. Durch die Erprobung diverser Verdünnungsstufen wurde deshalb empirisch die optimale Konzentration ermittelt.

Das gleiche Prinzip zeigt sich bei der Anzahl der durchgeführten Zyklen: Zu wenig Zyklen führen zu einer mäßigen Ausbeute, während eine zu hohe Anzahl den PCR-induzierten Bias verschärft und die Diversifikation der beobachteten Transkripte verringert (siehe S. 88 und Datson et al. 1999), sowie den Verlust des spezifischen 102bp Fragmentes zur Folge hat. Auch dieses Problem wurde empirisch gelöst.

Die Menge des resultierenden PCR Produktes gestaltet sich bezüglich der Nukleotidmenge dosis-abhängig. Nach dem Erreichen eines Maximums bei einer mittleren Konzentration sinkt bei weiter erhöhter Nukleotidkonzentration die Produktmenge wieder, da die Nukleotide vermutlich Mg^{2+} binden und so die Enzymaktivität negativ beeinflussen (Velculescu et al. 2000). Aus diesem Grund wird empfohlen, die Menge der Nukleotide zu titrieren. Der Versuch, die von verwendete Standardkonzentration von 1,5mM durch 0,4mM zu ersetzen, ergab ein deutlich schwächeres Resultat der PCR, so daß die ursprüngliche Konzentration beibehalten wurde.

4.2.1.3 Verdau der 102bp Fragmente mit NlaIII

Nach dem Abtrennen der Linker von den amplifizierten Ditags mittels NlaIII Verdau blieb bei der sich anschließenden Elektrophorese ein unverdauter Rest der 102bp Bande (Ditag mit zwei Linkern) zu sehen, was dreierlei Ursachen haben könnte.

1. Das Restriktionsenzym NlaIII verliert schnell seine Aktivität. Diesem Problem sollte dadurch begegnet werden, daß frisch gelieferte Chargen des Enzyms eingesetzt wurden.

2. Des weiteren weisen Velculescu et al. (2000) darauf hin, daß die während der Synthese des zweiten cDNS Stranges verwendete DNS Polymerase I Exonukleaseaktivität besitzt und die Persistenz dieses Enzyms später zu einem Abschneiden der die Erkennungssequenz für NlaIII enthaltenden Überhänge der Linker sowie der komplementären Überhänge der cDNS Fragmente führen könnte. Dies hätte zur Folge, daß die Linker über stumpfe Enden an die cDNS ligiert werden würden und die Ditag-Linkerkomplexe keine Erkennungssequenz für den zweiten Verdau mit NlaIII enthielten. Dies könnte die Unvollständigkeit dieses zweiten Verdaus verursachen. Denkbar wäre, eine zweifache Phenol-Chloroform-Extraktion im Anschluß an die Synthese der cDNS anstelle der durchgeführten einmaligen einzuführen, um so die Polymerase effektiver zu entfernen und damit zu einer besseren Ausbeute an Ditags (26bp Fragment) zu gelangen.

3. Angelastro et al. (2000b) zeigen, daß die Abtrennung der Linker von den Ditags durch die Einführung eines zusätzlichen Reinigungsschrittes vor dem enzymatischen Verdau der 102bp Fragmente wesentlich effizienter gestaltet werden kann. Die Autoren vermuten, daß

kontaminierende Substanzen, welche aus dem Polyacrylamidgel stammen, aus dem die Ditag-Linkerkomplexe aufgereinigt werden, die Aktivität von NlaIII verringern. Sie schlagen vor, zur entsprechend verbesserten Aufbereitung der 102bp Fragmente beispielsweise den *QIAquick Kit* (Qiagen) zu verwenden.

4.2.1.4 Behandlung der Ditag-Lösungen

Margulies et al. (2001) fanden in ihren SAGE Bibliotheken einen von der Behandlungsweise der Lösungen, welche freie Ditags enthielten, abhängigen Gehalt der Basen G und C. Dies hatte eine Verzerrung der Repräsentation der Tags entprechend ihrem Basengehalt zur Folge.
Um diesem Phänomen auf den Grund zu gehen, verglichen sie zwei SAGE Datenreihen, die von Material stammten, das bis zur PCR einem einzigen SAGE-Durchlauf angehörte, und das danach in zwei hinsichtlich der Behandlung der Lösungen mit freien Ditags unterschiedliche Gruppen geteilt wurde. Hierbei zeigte sich, daß bei Durchführung der Phenol-Chloroform-Extraktion der 26bp Fragmente bei Raumtemperatur der Anteil von G und C in den Tags durchschnittlich höher als 55% lag, während dieser in der Vergleichsgruppe, bei welcher die Extraktion auf Eis erfolgte, unter 50% war. Dieser zweite Wert entspricht Analysen der 3' '*untranslated regions*' (UTR), welchen die SAGE Tags entstammen. Deswegen gehen die Autoren davon aus, daß bei Raumtemperatur und außerdem niedriger Salzkonzentration eine Dissoziation der AT-reichen Tags auftritt, die in den Dokumentationen der Elektrophoresen nicht zu sehen ist. Durch die heterogene Komplexität der Lösung wird ein Reassoziation verhindert, so daß anschließend AT-reiche Tags in den Sequenzdaten unterrepräsentiert sind. Um dies zu umgehen, wird empfohlen, alle Schritte, die Lösungen freier Ditags involvieren, auf Eis und die Zentrifugationen bei 4°C durchzuführen. Velculescu et al. (2000) empfehlen außerdem die Verwendung von TE anstelle von LoTE, um so eine höhere Stabilität der freien Ditags zu gewährleisten. Eine Weiterentwicklung von SAGE namens LongSAGE (Margulies et al. 2001), welche mittels Verwendung eines anderen '*tagging*' Enzyms längere, das heißt stabilere, Tags synthetisiert, begegnet ebenfalls dieser Problematik.
Bei der Überprüfung der Daten der vorliegenden Arbeit fand sich in K1 ein durchschnittlicher GC-Gehalt pro Tag von 48%, in K2 von 53% (siehe auch Abb. 22). Diese Werte blieben nach der sekundären Elimination von Linkerartefakten konstant. Margulies et al. (2001) sehen SAGE Bibliotheken, die einen GC Anteil von mehr als 50% haben, als kritisch an. Bei einem Anteil von mehr als 55% soll eine eindeutige Verzerrung vorliegen. Es kann also davon ausgegangen werden, daß die der vorliegenden Arbeit zugrunde liegenden Daten akzeptabel sind. Dieses Ergebnis mag daran liegen, daß die Zentrifugationen der Lösungen mit freien

Ditags bei 4°C durchgeführt wurden.[8]

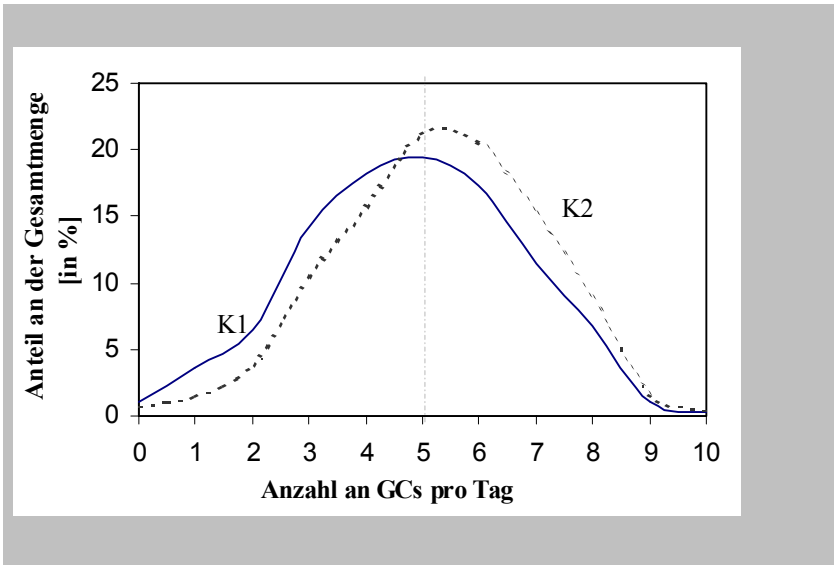

Abb. 22: Verteilung der SAGE Tags gemäß ihres GC-Gehaltes. Der Graph repräsentiert die Anteile der Tags der beiden Gruppen mit einem bestimmten GC-Gehalt (0-10 Basen). Um eine Verteilungskurve zu erhalten, wurde jeweils zwischen den einzelnen Punkten eine Linie gezogen. K2 weist im Vergleich zu K1 eine leichte Verschiebung nach rechts auf, was mit dem höheren Anteil an GC-reichen Tags in dieser Gruppe korrespondiert.

	K1	**K2**
GC	65325	73649
AT	70151	65501

Tabelle 10. Vierfeldertafel zur Überprüfung des CG Gehaltes. Angegeben ist die Anzahl der Basen.

Um festzustellen, ob der Unterschied im Basengehalt zwischen den beiden Gruppen statistisch signifikant ist, wurde ein Chi²-Test an der entsprechenden Vierfeldertafel (siehe Tab. 10) durchgeführt (H_0: Der Anteil der Basen GC ist in beiden Gruppen gleich. H_1: Der Anteil der Basen ist in beiden Gruppen verschieden; $\alpha = 0{,}05$; zweiseitiger Test). Bei einer ermittelten Wahrscheinlichkeit $p = 5{,}89*10^{-142}$ muß H_1 angenommen werden. Das heißt, daß der Gehalt an G und C zwischen den beiden Gruppen nicht als gleich erachtet werden kann. Dies ließe sich durch zufällige Schwankungen der Salzkonzentration der verwendeten Puffer oder durch Unterschiede in der Umgebungstemperatur oder der Temperatur der Lösungen bei der Verarbeitung der beiden Gruppen vermutlich nur teilweise erklären. Eine andere

[8] Eine SAGE Bibliothek, welche von adultem männlichen Gehirn (Maus) abstammt (Charst et al. 2000), hat einen GC-Gehalt von 54,5%, während zwei gekühlt erstellte Bibliotheken von Margulies et al. (2001) aus dem gleichen Gewebe eine Anteil von 48,3% beziehungsweise von 48,7% aufweisen.

Hypothese wäre folgende: Bei Betrachtung der beiden Linkersequenzen ergibt sich, daß diese eine sehr unterschiedliche Basenverteilung aufweisen: Einer der beiden Linker hat einen Anteil von 60% an G und C, während der andere einen von 30% besitzt. Dies führt zu der Hypothese, daß angesichts des hohen Anteils an Linkersequenzen Ditags, die als eine Tagkomponente eine GC-reiche Linkersequenz haben, dadurch stabilisiert werden und so einen höheren Anteil der Basen A und T zeigen. Eine Überprüfung der Linkerartefakte beider Gruppen ergab jedoch, daß K2, die Gruppe mit dem höheren GC-Gehalt, auch einen höheren Anteil an dem GC-reichen Linker besitzt (52,7% aller Linkerartefakte versus 43,7% in K1). Dies spiegelt sich leicht abgeschwächt in der Verteilung der Linkerbasen wider: die eliminierten Linkersequenzen von K2 haben einen GC-Gehalt von 46,9%, diejenigen von K1 von 45,0%, ohne jedoch diesbezüglich statistisch signifikant verschieden zu sein[9]. Dies bedeutet, daß die Hypothese, daß der Basengehalt der Linker den der Tags beeinflußt, nicht zutreffen kann.

Eine weiterer Erklärungsansatz für diesen Unterschied beruht auf der Tatsache, daß die Basen G und C im Vergleich zu den Basen A und T problematisch zu sequenzieren sind, weswegen sie sich tendenziell zu Beginn von Sequenzierungsläufen besser sequenzieren lassen, da hier die Qualität des Sequenzierungslaufes generell hochwertiger ist. Dies bedeutet, daß in einer Gruppe, in welcher die Sequenzierungsläufe tendenziell kürzer sind, mehr Gs und Cs enthalten sein müßten. Tatsächlich erwiesen sich in den Screening-PCRs die Inserts der Gruppe K2, welche den höheren Anteil an G und C besitzt, als kürzer im Vergleich zu Gruppe K1. Eine genauere Analyse dieser Hypothese müßte in weiteren SAGE Projekten untersucht werden.

4.2.1.5 Ligation der Ditags zu Polytags

Da die Anzahl der seriell hintereinandergeschalteten Tags wesentlich den Sequenzieraufwand und die Effizienz von SAGE bestimmt, trägt eine erfolgreiche Durchführung dieses Ligationsschritts entscheidend zur Senkung der Kosten pro SAGE-Tag bei. Die Länge der dabei entstehenden Ketten ist vor allem von zwei Parametern abhängig: der verwendeten Ditagmenge und deren Qualität.

---

[9] Um festzustellen, ob dieser Unterschied statistisch signifikant ist, wurde analog zum Gesamtvergleich ein Chi²-Test an der entsprechenden Vierfeldertafel (analog Tabelle 10) durchgeführt (H_o: Der Anteil der Basen GC ist in beiden Gruppen gleich. H_1: Der Anteil der Basen ist in beiden Gruppen verschieden; $\alpha = 0{,}05$). Bei einer ermittelten Wahrscheinlichkeit $p = 0{,}653$ kann die Alternativhypothese auf dem 5% Niveau nicht angenommen werden.

Um die zur Klonierung erforderliche Menge an DNS zu synthetisieren, wird im Protokoll Version 1.0c (Velculescu et al. 1997) die Durchführung von hundert bis zweihundert PCR Reaktionen (50µl) zur Amplifikation der Ditags empfohlen. Im Rahmen der vorliegenden Arbeit hingegen wurden parallel pro Kontrollgruppe fünfhundert Reaktionen durchgeführt, um diese DNS Menge bei möglichst geringer Anzahl von PCR Zyklen zu erreichen. Diese nebeneinander in großem Umfang laufenden Amplifikationen sowie eine äußerst sorgfältige Arbeitsweise ergaben qualitativ hochwertige Dimere. Dies zeigte sich darin, daß bei der Synthese der Konkatemere anstelle der im Protokoll angegebenen Ligationszeit (dreißig Minuten bis zwei Stunden) eine Inkubation von zehn Minuten ausreichte.
Die Qualität der eingesetzten Ditags wird außerdem durch ihren Reinheitsgrad nach der Gelreinigung der 26bp Fragmente bestimmt. Das Vorhandensein von angeschnittenen Linkern - trotz Gelreinigung - könnte beim Einbau dieser Fragmente einen vorzeitigen Kettenabbruch herbeiführen. Um eine sichere Auftrennung zu gewährleisten, wurde die Elektrophorese mit einer geringen Voltzahl, also sehr langsam, und geringen Ladungsmengen pro Bahn durchgeführt.
In der Literatur findet sich zu diesem Thema noch der folgende Vorschlag: Powell (1998) führt nach der Gelreinigung der 26bp Bande einen weiteren Reinigungsschritt ein. Sie zeigt, daß die abgeschnittenen Linker anhand im Amplifikationsschritt verwendeter biotinylierter Primer A und B nach dem Verdau mit NlaIII über die Bindung an magnetische Streptavidinpartikel effektiver entfernt werden können. Durch diese Modifikation des Originalprotokolls konnte die durchschnittliche Länge der klonierten Inserts von 620bp auf 740bp erhöht werden, was im Schnitt sechs Tags mehr pro Klon entspricht.
Da im Rahmen der vorliegenden Arbeit keine Schwierigkeiten bei der Synthese langer Polytags zu beobachten waren, scheint eine derartige Modifikation des Originalprotokolls in zukünftigen SAGE Durchläufen jedoch nicht notwendig zu sein.

4.2.1.6 Länge der klonierten Inserts

Allerdings ergab sich nach der Klonierung der Polytags in den Vektor das Problem, daß trotz der sehr effizienten Ligation per PCR lediglich relativ kurze Konkatemere in den Vektoren nachgewiesen werden konnten. In der Gruppe K1 wurden Ditagketten der Längen 700bp bis 2000bp in den Vektor ligiert, die durchgeführten Screening-PCRs ergaben jedoch nur eine mittlere Länge (ohne Vektorsequenz) von 500bp (±70bp). Bei K2 wurden sowohl Längen von 400bp bis 800bp als auch von 800bp bis 4000bp zur Klonierung verwendet. Dies führte zu einer durchschnittlichen Länge von 400bp (±100bp). Um die Leselängen der

Sequenzreaktionen (800bp) voll ausschöpfen zu können, wäre eine deutliche Verlängerung der Inserts wünschenswert. Auf diese Weise könnten nicht nur die hohen Sequenzierungskosten gesenkt, sondern auch den finanziellen Aufwand für die Reinigung der Plasmide halbiert werden.

Da es für kürzere DNS Fragmente wahrscheinlicher ist, daß sie in Vektoren ligiert werden, ist zu vermuten, daß vorrangig die jeweils kürzesten Ketten an erfolgreichen Reaktionen beteiligt waren. Es läge also nahe, in zukünftigen Experimenten schmalere Gelstreifen bei der Aufreinigung der Konkatemere zu schneiden, um so die Fragmente mechanisch in kleinere Längeneinheiten aufzutrennen, und diese dann separat in die Vektoren zu ligieren.

Ein weiterer Vorschlag zur Steigerung der Insertlänge findet sich bei Kenzelmann et al. (1999). Die Autoren weisen darauf hin, daß die während der Ligation der Ditags zu Polytags entstehenden kürzeren Ketten eventuell aggregieren und während der anschließenden präparativen Elektrophorese sozusagen versteckt mit längeren Fragmenten mitlaufen könnten. Diese Aggregation könnte durch Wasserstoffbrückenbildung zwischen den freien NlaIII Überhängen entstehen und durch die große Menge an Mg^{2+} aus dem Ligationspuffer stabilisiert werden (Kenzelmann et al. 1999). Diese Aggregate würden mit den langen Fragmenten zusammen aus dem Gel geschnitten und in den Vektor kloniert. Im Laufe dieses Geschehens würden die kurzen Fragmente desaggregieren und aufgrund ihrer Kürze bevorzugt in die Vektoren ligiert werden. Durch die Einführung eines Erhitzungsschrittes (65°C für 15 Minuten, dann für 10 Minuten auf Eis) vor der präparativen Gelelektrophorese der Polytags erreichte Kenzelmann eine Verlängerung der Ketten von durchschnittlich 22 Tags pro Klon auf 67. Die Autoren vermuten, daß durch die Erhitzung die Bindungen zwischen den Aggregaten aufgebrochen wurden und so keine Kontamination stattfinden konnte.

4.2.2 Probleme der Auswertung

4.2.2.1 Replikative Ditags und PCR Bias

Die mehrfach gezählten Ditagkombinationen resultieren - wenn Gründe wie fehlerhafte Eingabe in die Datenbank oder versehentliches mehrfaches Picken eines Klons ausgeschlossen werden - entweder aus einer nicht-repräsentativen Überamplifikation (Peters et al. 1999) und reflektieren somit einen PCR Bias (Velculescu et al. 1995) oder entstehen durch die Ligation sehr häufiger Tags (Welle et al. 1999), da diese eine hohe Wahrscheinlichkeit aufweisen, miteinander kombiniert zu werden. Um den PCR Bias nicht in die quantitative Auswertung von SAGE einfließen zu lassen, werden diese redundanten

Dimere vor der Datenauswertung eliminiert. Allerdings wird damit zugleich das Risiko eingegangen, die Verbindungen von sehr häufig vorkommenden Tags aus der Wertung überproportional auszuschließen. Welle et al. (1999) finden zum Beispiel für ein in ihrem Projekt sehr häufiges Gen wie dem, das die Cytooxygenase 2 kodiert, durch die Elimination der sich wiederholenden Ditagkombinationen eine Minderung der entsprechenden Tagmenge um 34% (von 1348 Tags auf 890 bei insgesamt 14000 ausgewerteten Tags). Dieses Problem, auf diese Weise die Häufigkeiten oft auftretender Gene zu unterschätzen, vergrößert sich prozentual, je mehr Tags insgesamt sequenziert werden. Aus diesem Grund werten die Autoren ihre Sequenzen in zwei getrennten Datenbanken aus und vereinen die ermittelten Tags und deren Häufigkeiten anschließend. Dieses Vorgehen ergibt der gewählte Versuchsaufbau zwangsläufig, so daß bei Verwendung der Daten der vorliegenden Arbeit (K1 plus K2) in späteren Vergleichen mit anderen Expressionsprofilen davon ausgegangen werden kann, daß diesem Fehler nach den entsprechenden Empfehlungen begegnet worden ist.

Bertelsen und Velculescu (1998) behaupten, daß die Kürze und Uniformität der SAGE Tags eine gleichmäßige Amplifikation der Ditags gewährleistet und so der Entstehung verzerrter Ergebnisse vorgebeugt wird. Spinella et al. (1999) zeigen jedoch mit einem Experiment, das die Ditag-PCR nachstellt, daß die Amplifikationsmenge sehr wohl von der Sequenz der Tags abhängig ist. Ditags, in welchen eines der beiden Tags eine zufällige Sequenz und das andere eine AT-reiche aufweist, lassen sich besser amplifizieren als Ditags mit dem gleichen zufälligen Tag und einem palindromischen Monotag (Spinella et al. 1999) Dies bedeutet, daß die Amplifikation des Ditags von den beiden Einzelbestandteilen beeinflußt wird, und daß non-palindromische Sequenzen im PCR Bias überrepräsentiert wären. Ob hier eine signifikante Beeinflussung der Resultate von SAGE besteht oder es sich um einen zu vernachlässigenden Effekt handelt, bleibt nachzuweisen.

Mit 5,3% ist der in den vorliegenden Daten detektierte Prozentsatz an replikativen Ditags im Vergleich mit Angaben aus der Literatur (Velculescu et al. (1995) 8,7% bei 1000 Tags, Velculescu et al. (1997) 8,34% von 68691 Tags) relativ niedrig. Dies könnte seine Ursache in der hohen Anzahl an parallel mit einer geringen Zyklenzahl durchgeführten PCRs - 500 gegenüber den 100 bis 200 im Protokoll angegebenen - haben. Der Vorteil einer solch niedrigen Rate an mehrfachen Ditags liegt darin, daß, um die angestrebte Menge an auswertbaren Tags zu erreichen, insgesamt weniger sequenziert werden muß und die Anzahl an verschiedenen Transkripten steigt (Datson et al. 1999), also die Diversität der erfaßten Stichprobe zunimmt.

4.2.2.2 Elimination der Linkersequenzen

Im Rahmen der Auswertung der Sequenzdaten mit der SAGE300 Software Version 3.00 ergab sich ein Anteil an Linkersequenzen von 11,05%. Bei genauerer Betrachtung dieser Artefakte zeigte sich, daß nicht nur diejenigen entfernt worden waren, die zu 100% den originalen Linkersequenzen entsprechen, sondern auch Tags, die um eine Base davon abweichen (siehe Tabelle 6). Allerdings waren hierbei nur Sequenzfehler im Sinne von Deletionen und Substitutionen beachtet worden. Um auch Linkertags, die Insertionsfehler aufweisen, berücksichtigen zu können, wurde im Rahmen der vorliegenden Arbeit ein Programm in der Programmiersprache C/C++ entwickelt. Dies durchsucht den gesamten Datensatz auf Sequenzen, die den Linkersequenzen unverfälscht entsprechen oder aber sich um eine Base davon unterscheiden. Dieser Ansatz ergab eine Linkerartefaktrate von 13,05%.
Problematisch an einer solchen Entfernung der Linkersequenzen ist die Tatsache, daß durch die Einbeziehung von Sequenzlesefehlern ein gewisser Unsicherheitsfaktor entsteht, bei welchen Tags es sich tatsächlich um Linkerartefakte handelt. Der hier (wie auch in der SAGE300 Software) gewählte Ansatz birgt einerseits die Gefahr, Transkripte zu entfernen, die den Linkern stark gleichen, so daß diese nie gemessen werden können, andererseits werden Linkerartefakte im Datensatz belassen, die um mehr als eine Base vom Original abweichen. Eine Klärung könnte nur ein (praktisch nicht umsetzbarer) Vergleich mit den Expressionsdaten anderer Methoden als externe Kontrolle erbringen. Aus diesem Grund sollte der methodisch-praktischen Bewältigung des Linkerproblems eine hohe Priorität zugewiesen werden (siehe S. 81).

4.2.2.3 Sequenzfehler

Fehler, welche die ursprüngliche Sequenz der Tags verändern, können das Ergebnis von SAGE nicht nur hinsichtlich der Entfernung von Linkerartefakten beeinflussen. Während die Häufigkeit eines Transkriptes durch einen Fehler bei der Sequenzierung des entsprechenden Tags verringert wird, wird diejenige des Trankriptes, welchem das Tag nun fälschlicherweise zugeordnet wird, erhöht. Bei Tags, welche oft vorkommen, spielt dieser zufällige Fehler nur eine geringe Rolle, wohingegen in dem Extremfall, daß ein Tag nur einmal vorhanden ist, dieses fälschlicherweise gar nicht detektiert wird. Es kann auch ein Tag etabliert werden, welches eigentlich gar nicht vorhanden ist. Aus diesem Grund wird auf der SAGE Internetseite (www. ncbi.nlm.nih.gov/SAGE, 5.12.2000) als Kompromißlösung vorgeschlagen, bei einer sequenzierten Tagmenge bis 250000 sämtliche Singletons zu eliminieren, bei Gesamttagmengen von mehr als einer Million auch Transkripte, welche zwei-

bis dreimal auftreten. Dieser radikale Ansatz wird von den Autoren selbst als nicht optimal bezeichnet.

Zhang et al. (1997) begegnen diesem Problem etwas differenzierter. Durch den Vergleich des bereits vollständig sequenzierten Hefegenoms mit einem SAGE Projekt, das das Expressionsprofil von Hefe untersuchte (Velculescu et al. 1995), wurde ein Sequenzierfehler von 0,7% pro Base ermittelt. Dies ergibt bei einer Taglänge von 10bp eine Fehlerrate von 6,8% (0,068 = 1 - $0{,}993^{10}$) je Tag[10]. Das heißt, daß die Wahrscheinlichkeit von 6,8% besteht, mindestens einen Fehler (bis zu zehn) in jedem Tag zu finden. Zur Korrektur wurde dieser Anteil an der Gesamtmenge der sequenzierten Tags von der Anzahl der detektierten unterschiedlichen Tags abgezogen, so daß sich die Anzahl der verschiedenen Transkripte reduzierte. Zhang et al. (1997) berechneten also 6,8% von 303.706 insgesamt sequenzierten Tags (entspricht 20652 Tags). Um diese Tagmenge wurden die 69393 verschiedenen Transkripte reduziert, so daß letztendlich lediglich 48741 unterschiedliche Tags ausgewertet wurden. Das Problem ist, daß auch hierbei die Auswahl der zu eliminierenden Gene sich nicht daran orientieren kann, wo der Sequenzfehler wirklich stattgefunden haben könnte.

Üblicherweise wird von einem Sequenzierfehler von 1% pro Base ausgegangen (Lal et al. 1999, Lash et al. 2000). Allerdings soll die Verwendung von Phred - einer Software zur Beurteilung der Sequenzierungsqualität - bei der Auswertung der Sequenzrohdaten die Fehlerrate um 40 bis 50% senken (Ewing et al. 1998), so daß ein 10bp langes Tag ungefähr eine Fehlerwahrscheinlichkeit (Summe über die Wahrscheinlichkeiten für genau einen Fehler bis genau zehn Fehler) von 5 bis 6% (1 - $0{,}995^{10}$ und 1 – $0{,}994^{10}$) aufweisen würde. An Zhangs Vorgehensweise orientiert, hieße das, daß in K1 bei 329 bis 394 und in K2 bei 314 bis 377 Singletons mindestens ein Sequenzfehler vermutet werden kann. Angesichts der Tatsache, daß die Häufigkeitskombination 1/1 oder 0/1 beziehungsweise 1/0 in dem vorliegenden Datensatz 11389-mal auftritt, erscheint es äußerst willkürlich, davon einen Bruchteil zu entfernen. Daraus läßt sich ablesen, daß es sinnvoll wäre, Kriterien zur systematischen Elimination und Sequenzfehlerkorrektur zu entwickeln. So wurde das bereits erwähnte Programm in C/C++ dazu verwendet, anhand der sequenzierten Linkerartefakte den in den Daten der vorliegenden Arbeit konkret vorhandenen Sequenzfehler abzuschätzen. Dazu wurden aus dem gesamten Datensatz diejenigen Tags ermittelt, welche sich in genau einer Base von den beiden Linkersequenzen unterscheiden. Die Anzahl wurde auf die Menge der insgesamt eliminierten Linkerartefakte, das heißt die exakt sequenzierten plus die um genau

[10] $0{,}993^{10}$ ist dabei die Wahrscheinlichkeit keinen Fehler zu finden.

eine Base abweichenden Tags, bezogen. Dies ergab eine Rate von 26,1%.[11] Dieser Wert gibt die Wahrscheinlichkeit wieder, daß sich in Tags *genau ein* Fehler befindet. Aus diesem empirisch ermittelten Schätzwert wurde basierend auf der Annahme, daß die Wahrscheinlichkeiten für *n* Fehler binomial verteilt sind, die Fehlerwahrscheinlichtkeit pro Base und anschließend für 0 bis 10 Fehler (siehe Tab. 11) errechnet.

Nach der Binomialverteilung ergibt sich für $p(a) = \frac{(1-a)^{(L-n)} \cdot a^n \cdot L!}{n! \cdot (L-n)!}$ mit *a* Wahrscheinlichkeit für den Fehler/Base, *L* der Taglänge und *n* Anzahl der Fehler im Tag. Es resultiert eine Fehlerrate pro Base von 3,65%. Auf dieser Basis listet Tabelle 11 sämtliche Wahrscheinlichkeiten für sämtliche mögliche Fehlerhäufigkeiten auf.

Fehler im Tag	0	**1**	2	3	4	5	6	7	8	9	10
Wahrschein-lichkeit *p(a)* [%]	69,0	**26,1**	4,44	0,45	0,03	$1{,}35 * 10^{-3}$	$4{,}25 * 10^{-5}$	$9{,}18 * 10^{-7}$	$1{,}30 * 10^{-9}$	$1{,}09 * 10^{-11}$	$4{,}14 * 10^{-13}$

Tabelle 11. **Wahrscheinlichkeiten für genau n Fehler**. Ausgegangen wird von einem 10er Tag und einem Fehler pro Base von 3,65%. Bei dem hervorgehobenen Wert handelt es sich um den aus dem Datensatz ermittelten Wert für genau einen Fehler/ Tag.

Die Wahrscheinlichkeit für Tags mit *mindestens* einem Fehler summiert sich zu 31,0% auf (Summer aller 10 Wahrscheinlichkeitswerte, *n* = 1 bis 10) - eine im Vergleich zu 5 bis 10% sehr hohe Fehlerwahrscheinlichkeit.

Um ergänzend eine Idee der Größe der Fehlerfortpflanzung zu vermitteln, sei folgendes betrachtet. Bei einer Änderung des empirischen Ausgangswertes (26,1% bei *n* =1) um 2 Prozentpunkte nach oben beziehungsweise unten ergibt sich für die summierte Fehlerwahrscheinlichkeit 34% beziehungsweise 28%, also keine wesentliche Größenänderung. Es ist wichtig zu beachten, daß es sich bei der Bestimmung des Sequenzierfehlers aus den aufgefundenen Linkerartefakten nicht um eine exakte Ermittlung des Fehlers handelt, sondern um eine annähernde Schätzung. Wie schon im Rahmen der Linkerelimination diskutiert, können zudem Transkripte, die den Linkern sehr ähnlich sind, die Fehlerrate fälschlicherweise erhöhen[12]. Auch beinhalten die Linkersequenzen Abfolgen

[11] Hierbei handelt es sich um einen geschätzten Wert, der den wahren Wert vermutlich übersteigt, da beispielsweise nicht alle um eine Base abweichenden Sequenzen Linkerartefakte sein müssen (siehe nächste Fußnote).

[12] Die Homologierecherche der Tags, die sich in einer Base von den Linkersequenzen unterscheiden (14.9.2002)

der Basen G und C, welche beim Sequenzieren besonders fehleranfällig sind. Es handelt sich bei diesem Schätzwert folglich um eine Obergrenze. Die Daten der auszuwertenden Tags liegen somit vermutlich unter dieser Fehlerangabe.

Vermieden werden derartige Probleme wie G oder C Abfolgen bei einer anderen Methode der Sequenzfehlerabschätzung. Piquemal et al. (2002) filtern aus den Sequenzdaten sämtliche Ditags heraus, die doppelt so lang sind, als es den Erwartungen entsprechen würde. Hierbei wird angenommen, daß der Verlust der Erkennungssequenz für NlaIII das Resultat eines Sequenzfehlers ist. Anhand der Anzahl dieser überlangen Ditags und der Fehler in der Nukleotidabfolge CATG läßt sich der Sequenzfehler abschätzen. Ein weiterer Ansatz zur Beurteilung des Sequenzfehlers findet sich bei Chen et al. (2002). Die Arbeitsgruppe setzt zwei künstlich generierte Oligonukleotide ein, welche im Aufbau SAGE Tags (inklusive Linker) entsprechen. In einem seperaten Experiment wird ein SAGE-Durchlauf vom Zeitpunkt der Ditagformierung imitiert. Die resultierenden Tagsequenzen werden im Vergleich mit den Originalsequenzen zu verwendet, um die Fehlerrate zu berrechnen. Dieser Ansatz hat den Nachteil, daß es sich nicht um eine interne Kontrolle handelt, sondern um ein weiteres Experiment, dessen Parameter möglicherweise nicht dem eigentlichen SAGE-Durchlauf entsprechen.

Um Sequenzfehler der in der vorliegenden Arbeit erstellten Expressionsprofilen näherungsweise, aber systematisch - im Gegensatz zu den bisher in der Literatur verwendeten unsystematischen Ansätzen (siehe oben) - zu korrigieren, wurde erneut das Programm in C/C++ eingesetzt. Diesmal wurden alle Singletons, die sich um eine Base von häufiger auftretenden Tags unterschieden, diesen zugeordnet, so daß deren Häufigkeit sich um ein Tag erhöhte, während das als fehlerhaft interpretierte Singleton aus dem Datensatz verschwand. Im Gegensatz zu den oben vorgestellten Ansätzen der Literatur versucht diese Vorgehensweise nicht nur durch Elimination die Anwesenheit potentiell nicht existenter Tags zu korrigieren, sondern auch den zweiten Effekt eines Sequenzfehlers - den Verlust von Tags - durch Zuordnung wieder wett zu machen.

Dennoch ist zu bedenken, daß dieser Ansatz nur solche Fehler zu korrigieren versuchen kann, die durch einen Irrtum in der Sequenz ein neues *einzelnes* Tag haben entstehen lassen. Tags dagegen, die aufgrund einer fehlerhaften Sequenz zufällig einem bereits vorhandenen Tag

---

ergibt 5 Tags (Masse: 14), die Genen zugeordnet werden können. Diese fünf Tags könnten also durchaus realen Transkripten entsprechen und würden nicht aus einem Sequenzierfehler resultieren. Dadurch verändert sich der geschätzte Sequenzfehler auf 3,59%/Base. Die summierte Wahrscheinlichkeit für Tags, mindestens eine fehlerhafte Base aufzuweisen, ist dann 30,7%. Dies wäre gegenüber 31% nur eine minimale Reduktion.

entsprechen und so diesem zugeordnet wurden, werden unkenntlich und damit einer Korrektur nicht zugänglich. Tags mit mehr als einer falsch gelesenen Base (laut Schätzung 5%, vergleiche Tabelle 11), entgehen dem hier gewählten Ansatz ebenfalls. Allerdings wurden bewußt lediglich Singletons zugeordnet, da andernfalls das Risiko zu sehr ansteigt, seltene Transkripte, die häufigen sehr ähneln (zum Beispiel durch Einzelnukleotidpolymorphismen), fälschlicherweise als Sequenzfehler aufzufassen. Zusammenfassend ist zu sagen, daß mit der hier vorgestellten Methode nur einem Bruchteil der möglicherweise vorhandenen Sequenzfehler begegnet werden konnte (8,4% versus maximal 31,0%).

Da es sich also bei allen vorgestellten Korrekturversuchen lediglich um Hilfskonstruktionen handelt, ist zu fordern, den Sequenzfehler, der die Ergebnisse von SAGE so deutlich beeinflußt (siehe S.167), möglichst gering zu halten. Zur Minimierung des Fehlers wäre für weiteren SAGE Projekten zu empfehlen, jedes Konkatemer doppelt, das heißt in zwei Richtungen zu sequenzieren. Allerdings müßte in einem solchen Fall eine Verdopplung der Kosten beziehungsweise eine Reduktion der sequenzierten Menge in Kauf genommen werden.

4.2.2.4 Homologierecherche

Mit einem 10bp langem SAGE Tag können über eine Million (4^{10}) verschiedener Transkripte unterschieden werden. Dennoch stellt eine 10bp lange Sequenz keineswegs eine perfekte Repräsentation eines Genproduktes dar (http://www.ncbi.nlm. nih.gov/SAGE, 5.12.00), so daß sich bei der Zuordnung der Tags zu bereits vorhandenen Gensequenzen trotz der Verwendung von UniGene Clustern, die speziell auf die Bedürfnisse von SAGE zugeschnitten worden sind (Lal et al. 1999, Lash et al. 2000, siehe auch S. 75), Uneindeutigkeiten ergeben können. Diese äußern sich darin, daß einem Tag mehrere Gene zugewiesen werden und umgekehrt.

Konstellation	Ursachen
ein SAGE Tag - mehrere Gene	Sequenzfehler der Datenbanksequenzen '*splitting*' von Sequenzen bei der Herstellung der UniGene Cluster zufällige Übereinstimmung von Transkripten verschiedener Gene beziehungsweise deren Varianten an der Stelle SAGE-Tags Verunreinigung der Datenbanksequenzen mit Vektorensequenzen
ein Gen - mehrere SAGE Tags	Spleißvarianten, die das 3' terminale Exon betreffen Einzelnukleotidpolymorphismen (SNPs) Auftreten multipler Polyadenylierungsstellen eines primären Transkriptes, die in einem Cluster zusammengefaßt werden '*lumping*' von Genen bei der UniGene Cluster Zusammenstellung
Tags ohne Zuordnung	Gen noch nicht bekannt oder sequenziert bisher unbekannte Spleißvarianten, die das 3' terminale Exon betreffen nicht im entsprechenden Cluster enthaltene SNPs Auftreten multipler Polyadenylierungsstellen eines primären Transkriptes, die in dem entsprechenden Cluster nicht enthalten sind inneres Oligo (dT) Priming bei der SAGE cDNS Synthese (innere Tags) unvollständiger zweiter Verdau mit NlaIII (innere Tags) Sequenzfehler des Tags SNP in der NLAIII Erkennungssequenz inneres Oligo (dT) Priming bei der cDNS Synthese der Datenbanksequenzen (3'Ende fehlt) Fehler der Datenbanksequenzen Verunreinigung der Datenbanksequenzen mit Vektorensequenzen falsche Ausrichtung klonierter Datenbanksequenzen vorzeitige Beendigung der Datenbanksequenz innerhalb des Transkriptes durch Endonukleaseverdau während des Klonierens
potentiell falsche Zuordnung	Sequenzfehler des Tags unvollständiger zweiter Verdau mit NlaIII (innere Tags) inneres Oligo (dT) Priming bei der SAGE cDNS Synthese (innere Tags) inneres Oligo (dT) Priming bei der cDNS Synthese der Datenbanksequenzen (falsches 3'Ende) Fehler der Datenbanksequenzen Verunreinigung der Datenbanksequenzen mit Vektorensequenzen falsche Ausrichtung klonierter Datenbanksequenzen vorzeitige Beendigung der Datenbanksequenz innerhalb des Transkriptes durch Endonukleaseverdau während des Klonierens

Tabelle 12. Gründe für uneindeutige oder möglicherweise falsche Resultate der Homologie-recherche unter Verwendung der speziell auf SAGE zugeschnittenen Datenbank (vergleiche www.ncbi.nlm.nih.gov /SAGE).

In den vorliegenden Daten trat das Problem, daß *einem* Tag *mehrere* Gene zugeordnet

werden, bei der exemplarischen Auswertung der fünfzig häufigsten Tags in 12% der Fälle auf. *Umgekehrt* ergab sich bei der Analyse der zuverlässig und eindeutig zugeordneten Gene (25 der 50) hinsichtlich weiterer Tags eine Tag-Cluster-Ratio von 2,64. Das heißt, daß im Mittel jedem dieser 25 Gene 2,64 Tags aus der Internetdatenbank zugeordnet werden konnte. Einige dieser zusätzlichen Tags fanden sich in den vorliegenden Daten wieder, so daß bei einer vollständigen Homologierecherche zum Beispiel Tubulin α (Mm. 196396) ein weiteres Tag (K1 und K2 je 10 mal) aufweisen würde.
Außerdem kann es Tags geben, für welche kein passendes Gen gefunden werden kann. Dies trat in 2% der betrachteten fünfzig Fälle auf.
Mögliche Ursachen für Probleme bei der Homologierecherche können Tabelle 12 entnommen werden. In den nachfolgenden Abschnitten werden einige dieser Problemfelder und die zugehörigen Lösungsansätze genauer betrachtet werden.

Fehler der Datenbanksequenzen

Um Sequenzfehlern in den veröffentlichten Sequenzen zu begegnen, wurde eine speziell entwickelte Datenbank ("*SAGEmap*") zur Homologiesuche verwendet, in welcher - wie unter 3.17.3 beschrieben - entsprechend dem erwarteten Sequenzfehler die untersten 10% der nach Rängen geordneten Tag-Cluster-Paare verworfen worden waren. Dennoch bleibt nach Angaben von Lash et al. (2000) ein Anteil von 13% unter allen möglichen Tags, die "*SAGEmap*" enthält, welchen mehr als ein Cluster zugeordnet wird. Dies hat seine Ursache darin, daß bisher nur ein geringer Anteil der in den Clustern vorhandenen Sequenzen von gut charakterisierten cDNS' abstammt (Mensch im Jahr 2000: 0,1%; Lash et al. 2000).
Um diese uneindeutigen Zuordnungen (ein SAGE-Tag: viele Gen-Cluster) zu untersuchen, könnte eine um ein oder zwei Basen erweiterte Analyse durchgeführt werden - so diese Basen überhaupt vorhanden sind, da dazu 11 oder 12bp lange Tags benötigt werden. Im Kontext der vorliegenden Arbeit erwies sich dieser Ansatz (für die beispielhaft betrachteten häufigen Transkripte) leider als wenig erfolgreich, da auch die weiter 3' liegenden Basen der in Frage kommenden Gene große Übereinstimmung aufwiesen. Dem Tag "CCTTTAATCC" beispielsweise ließen sich neun Gene zuordnen, die elfte Base C grenzte diese auf sieben ein. Um sämtliche Gene 100 % zu unterscheiden, wären in diesem Fall die Kenntnis von 35 Basen 3' der NlaIII Schnittstelle notwendig gewesen (Abb. 23), was den Rahmen von SAGE übersteigt.

```
Mm.1007:        tagtactcagaggcaggtggacctcttgagttcaaggctagtatt
        Mm.10305:       cagcacttgg
        Mm.104959:      tagcactcgggaggcagaggcaggcggatttctgagttcgaggccagcctggtctacagagtgagtt
        Mm.2024:        caggacttgggaggcagaggcaggca
        Mm.23692:
          cagcactcgacaggcaaaggccgtgagttgaaggccagtttggtctacacagcaagttccaggacagtcagcgc
        Mm.35814:       cagcactcgggaggcagaggcaggcagatttctgagttcgaggccagcctggtctaagagtgag
        Mm.3833:
          cagcactcaggagacagaggcaggcagaattctgagtttgaggccagcctggtctacaaagtgagttcc
        Mm.4784:
          cagcacttgggaggcagaggcagacagatttctaagttcaaggccagcctggtctacaaagtgagttcc
        Mm.56915:       cagcactcgggaggcagaggcaggtagatttctgagttcgaggccagcct
```

Abb. 23: 3' der 10. Base liegende Sequenzen der 9 reliablen Zuordnungen für "CCTTTAATCC".

In einem solchen Fall müssen alternative Möglichkeiten zur Identitätsbestimmung herangezogen werden. Es könnte zum Beispiel eine unabhängige Methode der Genexpressionsanalyse (zum Beispiel Northern Blot) angewendet werden, wie Lash et al. (2000) vorschlagen. Allerdings ist der erforderliche Arbeitsaufwand, Sonden und Hybridisierungen für sämtlich zur Diskussion stehende Gene anzufertigen, nur zu rechtfertigen, wenn es sich um eine geringe Anzahl von in Frage kommenden Genen handelt und das Tag aufgrund eines Vergleichprofils signifikant reguliert erscheint und somit ohnehin einer Validierung bedarf. Weitere Möglichkeiten der Identitätsbestimmung bestehen darin, eine PCR mit Tag spezifischen Primern durchzuführen ['*3' rapid amplification of cDNA ends*' (RACE) - PCR, welche die Sequenz des SAGE Tags als '*forward*' Primer benutzt, Michiels et al. 1999] oder das SAGE Tag als Oligonukleotidprobe zum Screening von cDNS Bibliotheken zu verwenden. Auf diese Weise kann diejenige cDNS gefunden werden, von welcher es abstammt. Diese kann dann weiter charakterisiert werden. Lee et al. (2002) entwarfen einen Ansatz, der die üblichen SAGE Tags in längere 3' ESTs umwandelt und so die Identitätsbestimmung erleichtert.

Tags, welchen kein Cluster zugeordnet werden konnte, können auf die gleiche Weise weiter untersucht werden. So können potentiell neue Gene schnell identifiziert werden beziehungsweise von solchen mit mangelnder Zuordnung aufgrund methodischer Schwächen (siehe Tabelle 12) unterschieden werden.

Fälschliche Zusammenstellung der UniGene - Cluster

Ein dem UniGene Projekt inhärentes Problem sind Fehler bei der Bildung der Cluster. Das heißt, daß Sequenzen, welche eigentlich zu verschiedenen Genen gehören, versehentlich in einem Cluster zusammengefaßt werden ('*lumping*') oder umgekehrt Sequenzen, die einem

Gen zuzuordnen sind, auf verschiedene Cluster verteilt werden ('*splitting*'). Ersteres erhöht die Tag - Cluster - Ratio, das heißt, daß einem Cluster verschiedene Tags zugeordnet werden, während zweiteres dafür verantwortlich sein kann, daß einem Tag mehrere Gene zugeschrieben werden. Laut Lash et al. (2000) liegt dieser Fehler jeweils unter 5%. Eine Kontrollmöglichkeit auf SAGE Ebene bietet sich hier nicht.

Variierende Transkriptsequenzen

Außerdem tragen, wie aus Tabelle 12 hervorgeht, variierende Sequenzen der Transkripte zu den Schwierigkeiten bei der Homologiesuche bei.
Einerseits können Variabilitäten in der Transkriptsequenz eines Gens Artefakte der Bibliothekssynthese darstellen. Dies ist der Fall bei Verunreinigungen der Datenbanksequenzen mit Vektorensequenzen, bei innerem Oligo(dT)-Priming während der cDNS Synthese der Datenbanksequenzen, bei falscher Ausrichtung der Inserts in Vektoren oder bei vorzeitiger Beendigung der Datenbanksequenz innerhalb des Transkriptes durch Endonukleaseverdau während des Klonierens. Lash et al. (2000) schätzen die Fehlerrate, die durch derartige Artefakte entsteht, auf 5 - 10%. Im Rahmen eines SAGE Projektes kann hierauf kein Einfluß genommen werden.
Andererseits können diese Variationen auch biologische Ursachen haben. Dies ist der Fall beim Auftreten von Spleißvarianten, die das terminale Exon eines Transkriptes betreffen, beim Auftreten von Einzelnukleotidpolymorphismen und von multiplen Polyadenylierungsstellen eines primären Transkriptes. Letzt genanntes Phänomen soll in 25% der UniGene Cluster vertreten sein (Lash et al. 2000).
Auf die Auswirkungen von Spleißvarianten und Einzelnukleotidpolymorphismen soll im folgenden genauer eingegangen werden.

a) Spleißvarianten

Dieser Aspekt ist nicht zu unterschätzen: Nach Mercante et al. (2001) sollen bis zu 35% der humanen Gene Spleißvarianten aufweisen.
Madden et al. (1997) detektieren beispielsweise für Cyclin G ein zusätzliches Tag, welches einer weiter 5' liegenden NlaIII Schnittstelle zuzuordnen ist. Dies führen sie darauf zurück, daß es von diesem Transkript eine noch unbekannte Spleißvariante gibt, da sich sonst keine Hinweise auf einen unvollständigen NlaIII Verdau (siehe unten) oder auf andere Ursachen für das Auftreten derartiger Tags ergeben hatten. Welle et al. (1999) fanden mittels BLAST Genbank Suche für eine mitochondriale 12S rRNS Tags von drei verschiedenen NlaIII

Erkennungssequenzen, was sie auf Längenunterschiede der Transkriptvarianten zurückführten.

Einen Hinweis auf das Vorhandensein von Spleißvarianten kann das Auftreten von inneren Tags darstellen, das heißt von Tags, die weiter 5' liegenden NlaIII Erkennungssequenzen entstammen. Um exemplarisch zu überprüfen, ob derartige Tags in den Daten der vorliegenden Arbeit vorhanden sein könnten, wurden fünf der häufigen und eindeutig zugeordneten Transkriptsequenzen untersucht. Dazu wurden sämtliche 3' aller NlaIII Erkennungssequenzen eines Transkriptes liegenden Nukleotidabfolgen (10bp) mit den Daten der vorliegenden Arbeit verglichen. Auf diese Weise fanden sich für das Myelinbasisprotein (GenBank Accession Nr. BC004704) und GAPDH (GenBank Accession Nr. NM_008084) jeweils ein potentielles inneres Tag (MBP: CCTTCTGTAG, GAPDH: TTTGTGATGG) von sehr geringer Häufigkeit (MBP: K1-0, K2-2; GAPDH: K1-0, K2-1). Im Gegensatz zu GAPDH sind in der Literatur für MBP verschiedene murine Spleißvarianten bekannt (de Ferra et al. 1985, Boccaccio et al. 1999), die unter anderem das terminale Exon involvieren, so daß bei dieser Boten - RNS die Detektion einer solchen Variante gut möglich sein kann. Im Fall von GAPDH sind andere Erklärungsversuche heranzuziehen (siehe unten). Interessanter Weise werden beide Tags eindeutig den jeweiligen Genen zugeordnet, was beim MBP bedeutet, daß die Spleißvariante in dem UniGene Cluster enthalten ist (als EST mit Poly(A)-Kennzeichnung und 3' Orientierungsangabe). Die Zuordnung im Falle von GAPDH bezieht sich dagegen auf ein EST ohne Poly(A)-Kennzeichnung, allerdings mit 3' Orientierungsangabe, so daß bei dieser Zuordnung von einer verkürzten Sequenz in der Datenbank ausgegangen werden kann. Zur sicheren Verifizierung solcher Spleißvarianten bleibt nur die Möglichkeit, zum Beispiel Northern Blots mit Sonden, welche alle Varianten detektieren können, durchzuführen und die Quantitäten zu vergleichen.

b) Einzelnukleotidpolymorphismen

Da Einzelnukleotidpolymorphismen ('*single nucleotid polymorphisms*', abgekürzt SNPs: Insertion, Deletion oder Substitution) sehr häufig sind, kann dieses Phänomen zum Beispiel dazu führen, daß - wenn wie im Fall der vorliegenden Arbeit die verwendete RNS von mehreren Individuen abstammt - für ein bestimmtes Gen unterschiedliche Tags entstehen, die unter Umständen nicht in den UniGene Clustern enthalten sind und somit nicht zugeordnet werden können. Wang et al. (1998) fanden bei einer Untersuchung von sieben Individuen alle

757bp einen solchen Polymorphismus. Umgerechnet[13] auf die verwendeten vier Mäuse gilt für ein 14bp langes SAGE Tag (inklusive NlaIII Schnittstelle), daß darin ein oder mehrere SNPs mit einer Wahrscheinlichkeit von circa 0,2% auftreten können. Eine derartige Variante könnte entweder den Austausch einer einzelnen Base in der Tagsequenz bewirken oder eine Änderung der gesamten Sequenz des Tags, wenn durch den Austausch der einen Base eine NlaIII Erkennungssequenz eliminiert oder neu eingeführt wird. Dies hätte zur Folge, daß das Tag nicht oder falsch zugeordnet werden würde.

Da ein bestimmter Polymorphismus in den meisten Fällen vermutlich nur bei einem der gepoolten Individuen auftreten würde, würde das betroffene Tag nur selten detektiert werden, so daß der auf diese Weise entstehende Fehler gering einzuschätzen ist. So entdeckten Welle et al. (1999) unter den 295 Genen, deren Tags sie mindestens 20 mal zählten, nur 4, die einen SNP aufwiesen (bei einem untersuchten Pool von 8 Individuen). Daraus ergibt sich eine Rate von 1,35%, die niedriger liegt als die von ihnen errechneten 2%. Die empirisch ermittelte Anzahl an Polymorphismen würde mit dieser erwarteten Rate vermutlich erst übereinstimmen, wenn auch die seltenen Tags in die Analyse einbezogen würden, da hier anteilmäßig mehr Tags mit SNPs zu finden wären.

Um derartige Sequenzvariation sicher von einem Sequenz- oder PCR-Fehler unterscheiden zu können, müßte jedoch die DNS der einzelnen Individuen untersucht werden - ein Aufwand, der dem möglichen Erkenntnisgewinn nicht entspricht. Es sei denn, alle gesicherten Polymorphismen, die für die Auswertung von SAGE Projekten von Relevanz wären, würden - wie von Baas und Tabak (1999) vorgeschlagen - in einer speziellen Datenbank gesammelt und somit die weitere Auswertung von SAGE Projekten erleichtern.

Unvollständiger erster NlaIII Verdau

Im Falle einer ineffizienten Durchführung des ersten NlaIII Verdaus oder bei unvollständiger Entfernung '*upstream*' liegender cDNS Fragmente nach diesem Verdau kann es dazu kommen, daß Tags nicht von der am meisten 3' liegenden Schnittstelle abstammen, sondern von einer weiter 5' liegenden Sequenz und somit inneren Tags entsprechen. Ein Problem der

---

[13] Nach der klassischen Populationsgenetik ist der Anteil der Polymorphismen proportional zu $(1^{-1} + 2^{-1} + 3^{-1} + ... + [n-1]^{-1})$, wobei n die Anzahl der untersuchten Genome darstellt. Demzufolge ergibt sich, wenn man von einer Rate von 1 SNP pro 757bp bei 7 untersuchten Individuen beziehungsweise einer Rate von 1/1159 bei 3 untersuchten Individuen ausgeht (Wang et al. 1998), für unseren Fall (4 Mäuse) eine Rate von 1/889 bis 1/903. Dies resultiert in eine Fehlerwahrscheinlichkeit von 0, 014% pro Base und damit für eine 14bp langes Tag von 0,2%.

Verwendung der speziell auf die Bedürfnisse von SAGE zugeschnittenen Datenbank liegt in diesem Fall darin, daß derartige Tags meistens entweder falsch oder fälschlicherweise gar nicht zugeordnet werden, da in der Datenbank UniGene Cluster ausschließlich aufgrund ihrer Sequenz 3' der am meisten 3' liegenden NlaIII Schnittstelle mit Tags gepaart werden. Da das Schneiden mit NlaIII - wie bei dem zweiten Verdau zu sehen ist - einen kritischen Punkt von SAGE darstellt, kann davon ausgegangen werden, daß dieses Problem nicht nur theoretischer Natur ist. Welle et al. (1999) finden unter den von ihnen detektierten Tags 5%, die von weiter 5' liegenden Schnittstellen abstammen. Allerdings können derartige Tags, wie bereits gesagt, auch dadurch entstehen, daß ein Transkript biologisch bedingt variable Längen aufweist.
Wenn ein auf diese Weise entstandenes inneres Tag zufälligerweise einem anderen Cluster zugeordnet werden kann, entgeht dieser Fehler der Aufmerksamkeit, solange keine erweiterte Analyse mit einer Datenbank durchführt wird, welche zusätzlich Tag-Cluster-Paarungen, die weiter 5' liegen, beinhaltet. In der vorliegenden Arbeit wurde eine derartige Untersuchung exemplarisch an fünf häufigen Genen durchgeführt (siehe oben: Spleißvarianten, S. 98). Dies ergab für zwei Gene potentielle innere Tags von sehr geringer Häufigkeit (nur in K2 detektiert, Gesamtmenge: 3 Tags).
Eine ebensolche beispielhafte Vorgehensweise wählen auch Chrast et al. (2000), um zumindest eine Idee von der Effizienz des Verdaus zu erhalten. Sie fanden im Vergleich zu den 3' liegenden SAGE Tags nur eine sehr geringe Anzahl an potentiellen inneren Tags: 1 bis 4 versus 68 bis 490. Daraus schließen die Autoren, daß von einem nahezu kompletten NlaIII Verdau ausgegangen werden kann. Auch die hier durchgeführte Analyse läßt mit der Einschränkung, daß es sich um keine vollständige Überprüfung handelt, diesen Schluß zu - zumal im Falle von MBP das innere Tag vermutlich einer Spleißvariante zuzuordnen ist (siehe oben). Ein singulär auftretendes Tag wie dasjenige, das GAPDH zugehörig zu sein scheint, ist außerdem verdächtig, nur aufgrund eines Sequenzfehlers existent zu sein. Es läßt sich allerdings keinem häufigerem Tag (größer oder gleich 2) - mit dem Unterschied von einer Base - zuordnen.
Eine punktuelle interne Untersuchung, wie sie hier geleistet wurde, sollte im Rahmen der Auswertung eines SAGE Projektes als minimale Kontrolle des Nla Verdaus erfolgen.

4.2.3 Implikationen für SAGE bei Sonderfällen der Boten-RNS

4.2.3.1 Transkripte ohne Erkennungssequenz für das Verankerungsenzym

Nicht von SAGE erfaßt werden können Boten-RNS Sequenzen, welche dem Verdau durch das verwendete Verankerungsenzym entgehen, da sie keine Schnittstelle dafür besitzen. Ein

Enzym wie NlaIII mit einer 4bp langen Erkennungssequenz schneidet durchschnittlich alle 256bp (4^4), wobei die Mehrzahl der Boten-RNS' beträchtlich länger sind (Velculescu et al. 1995). Kürzere Transkripte haben dennoch generell eine höhere Wahrscheinlichkeit dafür keine Erkennungssequenz zu besitzen: Wenn Sequenzen gänzlich per Zufall erstellt werden würden, hätte ein 2kb langes Transkript eine Chance von größer 99,9% mindestens eine NlaIII Schnittstelle aufzuweisen, ein 1kb langes eine 98%ige Chance, ein 0,5kb langes eine 85%ige Chance und ein 0,25kb langes eine Chance von 62% (Welle et al. 1999). Beispiele für Transkripte ohne Schnittstelle finden sich in der Literatur einige. Welle et al. (1999) können mit SAGE drei der in anderen - ihrem Projekt entsprechenden - cDNS Bibliotheken häufigsten Transkripte nicht detektieren, was sich bei zwei Transkripten (Cytochrom C Oxidase 7a, 341bp lang, und ribosomales Protein S21, 343bp lang) auf die fehlende NlaIII Schnittstelle zurückführen läßt. Kal et al. (1999) berichten von einem Transkript in der Hefe (TPI 1 Gen), auf welches dies ebenso zutrifft. Um dieses Phänomen exemplarisch für die vorliegende Arbeit zu überprüfen, wurden Sequenzen einiger häufiger Transkripte in murinen UniGene Gehirnbibliotheken aus EST Sequenzierungsprojekten (Lib. 16, 200, 205, 230, 483, 264, 280 und 161 - www.ncbi.nlm.nih.gov/ UniGene, 10.10.01), deren potentielle Tags in den hier vorliegenden Daten nicht wiederzufinden waren, hinsichtlich der NlaIII Erkennungssequenz analysiert. Diese war jedoch in sämtlichen Sequenzen vorhanden.
Die Voraussetzung um fehlende Schnittstellen zu erkennen, ist also, daß zu dem untersuchten Gewebe externe Kontrollmöglichkeiten existieren. Das heißt, daß (quantitative) Expressionsdaten, welche mit einer anderen Methode erstellt wurden, vorhanden sein müssen. Dies ist eine Bedingung, die nicht immer gegeben ist. Eine interne Kontrolle wäre, weitere SAGE Durchläufe mit anderen Verankerungsenzymen durchzuführen - ein Ansatz, bei dem jedoch die hohen Kosten, der dafür notwendige Zeitaufwand und unter Umständen die erforderliche doppelte Menge an Ausgangsmaterial zu berücksichtigen ist.

4.2.3.2 Besondere Lage der Erkennungssequenz für NlaIII (5')

Theoretisch könnten Boten-RNS Moleküle, deren am meisten 3' liegende Schnittstelle für NlaIII sich überdurchschnittlich weit 5' befindet, unterrepräsentiert sein, da während der cDNS Synthese der erste Strang unter Umständen nur ungenügend verlängert wird. Das würde bedeuten, daß in solchen Fällen die cDNS Synthese zu früh abgebrochen wird, um die NlaIII Erkennungssequenz zu erreichen, und es in der Folge dem cDNS Fragment an dieser mangelt. Welle et al. (2000) schätzen, daß dieses Problem ab einem Abstand der Erkennungssequenz vom Poly(A)-Schwanz von mehr als 500bp relevant sein könnte. Die

Wahrscheinlichkeit, daß über eine solche Länge keine NlaIII Schnittstelle zu finden ist, liegt bei 15%.

Beim Vergleich zweier SAGE-Expressionsprofile fanden Welle et al. (2000) ein sachlich unlogisches Expressionsmuster eines Gens, welches sich nach Überprüfung anhand einer RT-PCR als quantitativ falsch herausstellte. Als Erklärung wird von den Autoren angeführt, daß die cDNS Synthesen der beiden SAGE Durchläufe, welche miteinander verglichen wurden, im Gegensatz zu denjenigen der RT-PCR nicht parallel durchgeführt wurden. Sie könnten folglich eine unterschiedliche Effizienz aufweisen, was unterschiedliche Längen der cDNS Fragmente zur Folge hätte. Ein Vergleich der Quantitäten des 5' und des 3' Endes des cDNS Fragmentes für das betroffene Transkript aus einer nicht weiterverarbeiteten Charge der SAGE cDNS Synthese bestätigte die Vermutung.

Dies zeigt, daß auch hier nur externe oder interne mit einem zweiten Verankerungsenzym (siehe oben) Kontrollen eventuell vorhandene Probleme aufzeigen können. So ergab für die vorliegende Arbeit eine erneute Untersuchung der in anderen murinen zerebralen cDNS-Bibliotheken (siehe vorherigen Abschnitt) häufigen und in den hier vorliegenden Daten nicht oder nur in sehr geringem Maße vorhandenen Transkripte beim "CUG RNA bindenden Protein 2" (GenBank Accession Nr. NM_010160), daß die Erkennungssequenz für NlaIII mehr als 500bp 5' des Poly(A)-Schwanzes liegt. Dies könnte eine Ursache dafür sein, daß dieses Transkript in den vorliegenden Daten nicht vorhanden ist. Allerdings ließ sich ein Kontrollgen (AKAP) mit einer Länge von 4,2kb ohne Degenerationszeichen im cDNS Southern Blot nachweisen. Dies spricht gegen eine derartige Ineffizienz der cDNS Synthese.

4.2.3.3 Besondere Lage der Erkennungssequenz für NlaIII (3')

Die Wahrscheinlichkeit, daß sich eine NlaIII Schnittstelle innerhalb von 10bp 5' des Poly(A)-Schwanzes befindet, liegt bei 4%. Daraus resultierende SAGE Tags hätten 3' mindestens 4 As. Diese Konstellation trifft bei den Daten der vorliegenden Arbeit auf 281 Tags zu (Masse: 716), was knapp 2% aller Tags des Projektes entspricht und somit unterhalb des erwarteten Rahmens liegt. Wenn die Schnittstelle des Verankerungsenzym unmittelbar an den Poly(A)-Schwanz grenzt, besteht das korrespondierende Tag sogar nur aus der Base A und ist nicht mehr eindeutig zuzuordnen. In Fall der vorliegenden Arbeit ergaben sich für dieses Poly-A-Tag, das hier 0,42% der Masse aller Tags ausmacht, 27 mögliche Genzuschreibungen. Das Problem hierbei ist, daß diese Vielzahl nur schwierig weiter eingegrenzt werden kann, da weder eine erweiterte Recherche mit einem 11 oder 12bp langen Tag durchgeführt werden kann - es kämen nur weitere As hinzu - noch dieses Tag als Probe zur Druchführung eines

Screenings einer cDNS Bibliothek oder als Vorwärtsprimer in einer PCR sinnvoll einsetzen kann. Es bliebe die Möglichkeit, anhand einer anderen Methode zur quantitativen Messung der Genexpression das Expressionsmuster aller in Frage kommenden Gene mit dem per SAGE ermittelten zu vergleichen - bei fast dreißig Genen ein großer Aufwand, der ökonomisch nur tragbar wäre, wenn es sich nach Vergleich zweier Profile um ein sehr interessantes Transkript handeln würde. Einige Autoren (zum Beispiel Welle et al. 1999) verwerfen dieses Tag deswegen und schließen es von der Auswertung damit vollständig aus. In der Analyse der vorliegenden Arbeit wurde es beibehalten, da es hier vorrangig um die exemplarische Etablierung von SAGE ging und nicht um eine exakte Identifizierung der Zugehörigkeit aller Transkripte.

4.2.3.4 Transkripte, deren komplementärer Strang die Enzymerkennungssequenz für BsmFI enthält

Wenn der dem SAGE Tag komplementäre Strang nach dem ersten NlaIII Verdau innerhalb von circa 20bp 3' des Linkers in Gegenrichtung eine Erkennungssequenz für BsmFI enthält (entspricht der Sequenz GTCCC im eigentlichen Strang), könnte eine Unterrepräsentation dieses Transkriptes resultieren, da der Verdau mit der Typ II S Endonuklease das Tag verkürzen oder von dem Linker abschneiden könnte. Welle et al. (1999) schätzen, daß dieses Problem 2% der cDNS' betreffen könnte. Madden et al. (1997) berichten von einem murinen Transkript ($p21^{WAF1/CIP1}$), das aus diesem Grund um 80% weniger, als nach Literaturangaben zu erwarten gewesen wäre, detektiert wurde.

Die Überprüfung der verlängerten Sequenzen der eindeutig zugeordneten unter den fünfzig häufigsten Tags der vorliegenden Daten ergab ein Transkript (Uba52, GenBank Accession Nr. NM_019883), dessen Messung auf diese Weise beeinflußt worden sein könnte (catgtgaccccgggaccaaataaagtccc). Jedoch erst eine unabhängige Quantifizierung dieser Boten-RNS oder ein zweiter - ökonomisch nicht vertretbarer - SAGE-Durchlauf, in welchem eine andere Typ IIS Endonuklease verwendet werden würde, könnte diese Hypothese bestätigen oder widerlegen.

Bei der Kontrolle der Sequenzen sämtlicher Tags zeigte sich, daß 82 Tags (entspricht 0,58% von 14159 verschiedenen Transkripten) die potentiell problematische Nukleotidabfolge gtccc enthielten. Dies korrelierte mit einer Menge von 211 Tags (0,77% von 27499 Tags insgesamt). Auch hier läßt sich nur vermuten, daß die Messung des Expressionsniveaus der zugehörigen Gene nicht korrekt ist. Eine Klärung der Verhältnisse ließe sich nur durch oben genannte Vorgehensweisen vollbringen.

Um auf der anderen Seite exemplarisch festzustellen, ob es Transkripte gibt, die in den hier vorliegenden Daten aufgrund dieser Problematik nicht oder kaum detektiert wurden, wurden die bereits erwähnten Sequenzen anderer Bibliotheken darauf überprüft, ob sie die Sequenz gtccc innerhalb von ungefähr 20bp 3' der ersten NlaIII Erkennungssequenz enthalten. Dies ergab bei einem Transkript (Enolase 2) einen entsprechenden Hinweis darauf, weshalb es im Rahmen der vorliegenden Arbeit möglicherweise nicht (K1) beziehungsweise kaum (K2: einmal) detektiert worden war (catgtcccacagtt). Auch hier gilt, daß eine zweite Methode oder eine Methodenänderung zur Veri- oder Falsifizierung dieser Hypothese notwendig wäre.
Des weiteren wäre es sinnvoll, wenn Software, welche SAGE Daten inhaltlich analysiert, Tagsequenzen, welche selbst schon derartige problematische Sequenzen enthalten, mit einer Warnung versehen würde. Das gleiche gilt für die Zuordnung von Datenbanksequenzen zu den detektierten Tags. Auch hier wären automatische Hinweise auf Sequenzanteile, welche gtccc entsprechen, zweckmäßig.

4.2.4 Beurteilung des quantitativen Resultats

4.2.4.1 Verteilung der Häufigkeiten

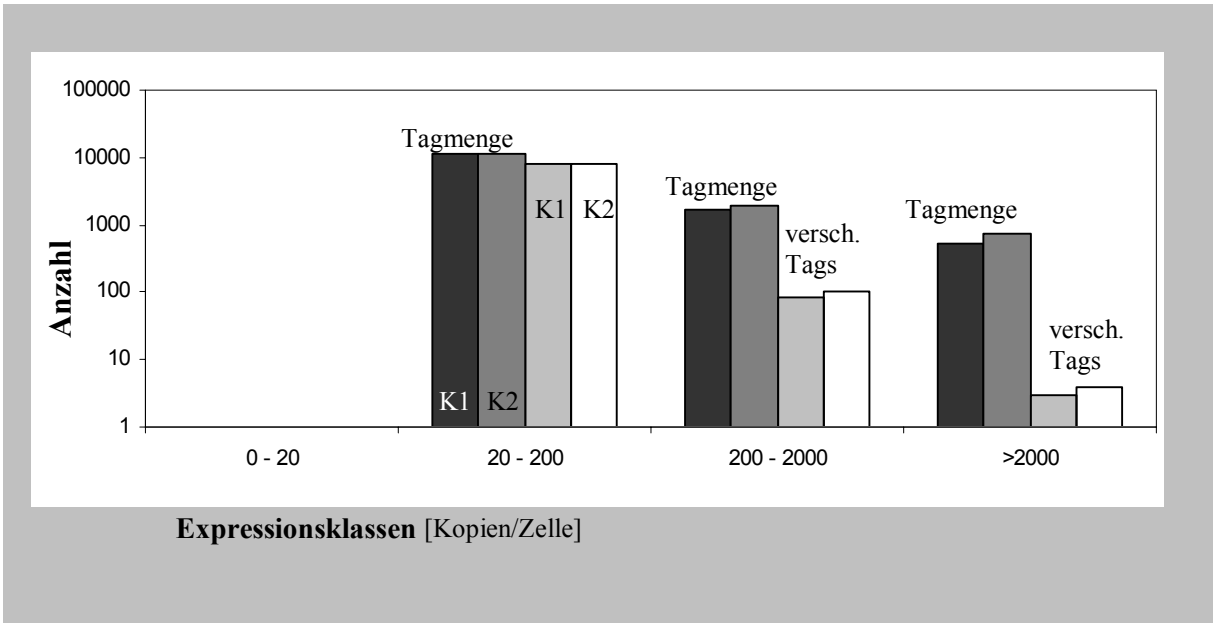

Abb. 24: Vergleich der Häufigkeiten und der Anzahl der verschiedenen Tags in den verschiedenen Expressionsklassen. Die Bezeichnung "Tagmenge" bezieht sich auf die Gesamt-anzahl der jeweils detektierten Tags, "versch. Tags" auf die Anzahl der verschiedenen Tags. Über die kleinste Expressionsklasse (null bis zwanzig Transkripte) ist keine Aussage möglich (siehe S. 79), da im vorliegendem Datensatz 1 Tag circa 20 Kopien/Zelle entspricht, so daß hier eine Leerstelle bleiben muß.

Wie unter Punkt 4.1.4.18 dargestellt befinden sich 99% der detektierten unterschiedlichen Tags in der niedrigsten Expressionsklasse (kleiner 200 Kopien/Zelle, entsprechend kleiner 10 Tags). Die Umrechnung der Taghäufigkeiten in Anzahl der Transkripte pro Zelle basiert auf der Annahme, daß 300000 Transkripte in jeder Zelle existieren (Hastie und Bishop 1976). Wenn die Häufigkeit der Tags der niedrigsten Expressionsklasse betrachtet wird, entsprechen sie 82% der Menge aller auswertbaren Tags (siehe auch Abb. 24, "Anzahl" logarithmisch aufgetragen).
Dieser Befund steht im Gegensatz zu dem, was zum Beispiel Zhang et al. (1997) über Untersuchungen an kolorektalem Gewebe berichten. Dort synthetisiert zwar ebenso die Mehrzahl der Gene nur wenige Transkripte, die Menge dieser Transkripte stellt jedoch lediglich 25% der gesamten Boten-RNS Masse dar. Das in den in dieser Arbeit vorliegenden Daten beobachtete Verteilungsmuster spiegelt also die Komplexität der zerebralen Genexpression wieder. Dies deckt sich mit den Ergebnissen von Velculescu et al. (1999b), welche in einer Metaanalyse von 84 humanen SAGE Bibliotheken die Transkriptverteilung des Gehirns als die komplexeste aller Gewebe bezeichnen.

4.2.4.2 Repräsentativität

Da SAGE zum Ziel hat, Transkriptome umfassend zu untersuchen, ist die Repräsentativität der Ergebnisse ein wichtiger Aspekt dieser Methode. SAGE quantifiziert Genexpression zwar absolut, führt dies jedoch an Zufallsstichproben durch. Die Frage nach der Repräsentativität der Resultate eines SAGE-Durchlaufes ist demnach nur durch die Berechnung von Wahrscheinlichkeiten zu beantworten. Im Folgenden soll dieser Frage in bezug auf die beiden Profile der vorliegenden Arbeit nachgegangen werden.
Den Berechnungen, wie wahrscheilich es ist, eine Boten-RNS eines bestimmten Expressionsniveaus mindestens einmal zu detektieren, wurde eine Grundpopulation N von 300000 Boten-RNS Molekülen pro Zelle (Hastie und Bishop 1976) und die Binomialverteilung für die Gegenwahrscheinlichkeit, die entsprechende RNS nicht zu ziehen, zugrunde gelegt:

$$p = 1 - \binom{N-m}{n} \Big/ \binom{N}{n} \quad \text{(Gleichung 1).}$$

Die Ergebnisse sind **Tabelle 13** zu entnehmen.

Expressionsniveau *m* [Kopien/Zelle]	Wahrscheinlichkeit *p*, ein Transkript *mindestens einmal* zu detektieren bei einer Stichprobe *n* von:	
	14000 (K1 *oder* K2)	28000 (K1 + K2)
1	4,7%	9,3%
10	37,9%	62,5%
25	69,7%	91,4%
50	90,8%	99,3%
100	99,2%	99,9%
1000	100%	100%

Tabelle 13. Wahrscheinlichkeiten, ein Transkript *mindestens einmal* zu detektieren. Es sind für verschiedene Expressionsniveaus beispielhaft die gerundeten Wahrscheinlichkeiten der beiden Expressionsprofile einzeln beziehungsweise zusammengenommen dargestellt.

Es ist zu beachten, daß im Gehirn bei der Mehrzahl der Gene die Expression relativ niedrig ist. Das heißt, daß diese Gene, welche sehr wenig, kurz oder nur in einem Bruchteil der untersuchten Zellpopulation exprimiert werden, in der vorliegenden Untersuchung nur mit einer geringen Wahrscheinlichkeit detektiert werden. Daraus folgt, daß die Abwesenheit eines Gens in den vorgestellten Expressionsprofilen nicht zwangsläufig bedeutet, daß dieses Gen in dem untersuchten Gewebe nicht exprimiert wird, oder daß es aufgrund der oben diskutierten möglichen Einschränkungen nicht erfaßt werden kann. Die dargestellten Wahrscheinlichkeitswerte weisen auf die Begrenzungen der Repräsentativität der vorliegenden Untersuchung hin. Wie bereits in der Einleitung (siehe S. 22) dargestellt, ist die Frage nach der Reliabilität von SAGE eng mit dem Thema Repräsentativität verknüpft, weswegen weiter unten nochmals auf dieses Thema eingegangen wird (S. 139ff).

4.3 Fazit der Praxis von SAGE

SAGE hat als digitales und offenes Verfahren ein großes Potential. Dies zeigt sich auch in der wachsenden Anzahl an Publikationen von SAGE Projekten (Patino et al. 2002). Es benötigt vor der Durchführung der Genexpressionsanalyse keinerlei Information über die Sequenzen der Transkripte und ihren biologischen Hintergrund und nimmt eine direkte Quantifizierung der Genexpression vor. Durch das serielle Koppeln der SAGE-Tags wird ein hoher Durchlauf gewährleistet, was die Methode sehr effizient macht. Um dieses Leistungsvermögen optimal zu entfalten, wird SAGE ständig weiterentwickelt. Im Folgenden sollen die bei der Etablierung der Methode im Kontext der vorliegenden Arbeit deutlich gewordenen Probleme und deren Lösungsmöglichkeit zusammengefaßt und gewinnbringende Modifikationen herausgestellt werden.

Auf der Seite der molekular-biologischen Durchführung stehen dabei im Vordergrund:

- Die Kontamination mit Linkersequenzen senkt die Zahl der auswertbaren Tags und somit die Effizienz und die Möglichkeiten eines SAGE Projektes. Da der Anteil an Linker(artefakten) in den Daten der vorliegenden Arbeit überdurchschnittlich hoch ist, wäre bei einem erneuten SAGE- Durchgang auf die Umsetzung der diskutierten Vorschläge zur Lösung dieses Problems ganz besonderen Wert zu legen.
- Die Begradigung der Tags nach dem BsmFI Verdau mit Klenow erfolgt bei 11°C statt 37°C wie empfohlen, um die Exonukleaseaktivität des Enzyms zu verringern und damit die Wahrscheinlichkeit für zu kurze Tags zu senken.
- Der Restriktionsenzymverdau der 102bp Fragmente mit NlaIII verläuft nicht vollständig, was die Effizienz von SAGE limitiert. Hier wäre es sinnvoll, zusätzliche Reinigungsschritte einzuführen.
- Die Behandlung der Lösungen, welche freie Ditags enthalten, beeinflußt die Stabilität der Tags und entscheidet über deren letztendlichen Gehalt an den Basen G und C. Die beiden parallel verarbeiteten Gruppen K1 und K2 weisen hier einen signifikanten Unterschied auf, über dessen Ursache nur Vermutungen geäußert werden können, die weiterer Untersuchungen bedürfen.
- Die Durchführung von 500 parallelen PCRs mit einer relativ geringen Zyklenzahl - statt 100 wie empfohlen - ergibt qualitativ hochwertige Amplifikate bei erhaltener Diversität. Dies zeigt sich in der geringen Anzahl von doppelten Ditags.
- Bei der Ligation der Ditags zu Konkatemeren ist aufgrund der hohen Qualität der Ditags (hohe Anzahl von parallel durchgeführten PCRs und langsame Elektrophorese mit

niedriger Spannung) eine überdurchschnittliche Effizienz zu beobachten, so daß es in diesem Fall nicht erforderlich scheint, die Modifikationen aus der Literatur einzuführen. Außerdem genügt eine deutlich verkürzte Inkubationszeit.

- Aufgrund der Tatsache, daß tendenziell kurze Konkatemere in die Vektoren ligiert wurden, wäre zu empfehlen, nach der Elekrotphorese sehr schmale Gelstreifen zu wählen und die aufbereiteten Fragmente getrennt in die Vektoren zu ligieren.

Im Rahmen der Auswertung der Sequenzdaten sind folgende Punkte zu beachten:

- Um den von der PCR-induzierten Bias, der das quantitative Resultat von SAGE verzerrt, auszuschließen, werden redundante Ditags eliminiert. Deren Rate ist in den Daten der vorliegenden Arbeit vergleichsweise niedrig, was den Vorteil hat, daß der Anteil an unnötig sequenzierten Tags, welche die Effizienz des Projektes senken und die relativen Kosten erhöhen, gering ist. Hier mögen die in sehr hoher Anzahl parallel mit niedriger Zyklenzahl durchgeführten PCRs von Vorteil sein - eine Vorgehensweise, die weiteren SAGE Projekten zu empfehlen ist.
- Problematisch gestaltet sich die Elimination der Linkersequenzen, was die Notwendigkeit einer methodischen Lösung (siehe oben) dieses Problems betont. Nachdem die SAGE300 Software Version 3.00 hier Mängel aufweist, wurden mittels eines zusätzlich geschriebenen Computerprogramms sämtliche Tags aussortiert, welche den originalen Linkersequenzen entsprechen beziehungsweise um eine Base von diesen abweichen. Es muß diesbezüglich beachtet werden, daß die Art und die Kriterien der Linkerelimination in Publikationen zu SAGE nicht erwähnt werden.
- Ganz entscheidend wird das quantitative Ergebnis eines SAGE Projektes durch den Sequenzfehler beeinflußt, der in der vorliegenden Arbeit als Schätzwert aus den Linkerartefakten über dem erwarteten Wert liegt. Zur nachträglichen Korrektur dieses Fehlers existieren verschiedene, allesamt wenig zufriedenstellende Ansätze, die sich nur auf die unsystematische Entfernung einmal auftretender Tags konzentrieren. Um potentiell fehlerhafte Tags (Singletons) nicht nur gezielt zu entfernen, sondern auch ihren häufiger auftretenden Partnern, von welchen sie vermutlich abstammen, zuzuordnen, wurde ein Computerprogramm entwickelt. Da auch diese Vorgehensweise nur einen Teil des Sequenzfehlers abdeckt und eine Hilfskonstruktion darstellt, ist für zukünftige SAGE Projekte unbedingt zu diskutieren, den Fehler durch Sequenzierung in beiden Richtungen zu minimieren. Sinnvoll wäre auch, um eine Idee von der Größe des Sequenzfehlers des jeweiligen Projektes zu erhalten, diesen zum Beispiel aus Linkerartefakten wie in der

vorliegenden Arbeit abzuschätzen.

- Die Homologierecherche stellt einen Bereich dar, der mit vielfältigen Schwierigkeiten behaftet ist. Tags können mehreren Genen, ein Gen kann mehreren Tags, Tags können falsch oder gar nicht zugeordnet werden. Die Ursachen hierfür liegen einerseits bei den Gendatenbanken und sind zum anderen Teil SAGE spezifisch. Darunter fallen zum Beispiel Spleißvarianten, Einzelnukleotidpolymorphismen und ein unvollständiger erster Verdau mit NlaIII. Diese Probleme lassen sich - wenn überhaupt - nur durch Erweiterungen von SAGE oder Anwendung alternativer Methoden wie Northern Blot beheben. Um die Güte des Verdaus mit NlaIII exemplarisch zu überprüfen, kann eine interne Kotrolle der Daten auf innere Tags, also Tags, die von weiter 5' liegenden Schnittstellen stammen, erfolgen. In den vorliegenden Daten ergab sich dabei kein negativer Hinweis.

Des weiteren wird SAGE durch Sonderfälle der Boten-RNS Gestalt beeinflußt:

- Transkripte ohne Erkennungssequenz für das Verankerungsenzym sind SAGE nicht zugänglich. Bei einem Enzym wie NlaIII, das alle 256bp schneidet, tritt dieses Problem jedoch nur selten auf. Eine stichprobenartige Untersuchung der hier vorliegenden Daten ergab kein Transkript mit einer derartigen Konstellation.
- Boten-RNS Moleküle, bei welchen die erste Erkennungssequenz für NlaIII sehr weit 5' liegt, können aufgrund ungenügender Verlängerung des ersten Stranges während der cDNS Synthese unterrepräsentiert sein. In der durchgeführten exemplarischen Analyse fand sich ein einziges Transkript, das hiervon möglicherweise beeinflußt sein könnte.
- Auch Transkripte, deren komplementärer Strang innerhalb von circa 20bp 3' des Linkers in Gegenrichtung die Erkennungssequenz von BsmFI enthält, laufen Gefahr, unterrepräsentiert zu sein, da der Verdau mit dem Enzym diese verkürzen oder vom Linker abschneiden könnte. Hier fand sich ebenfalls nur ein Transkript unter den häufigsten fünfzig, auf das diese Möglichkeit zutreffen könnte. Weniger als 1% aller Tags enthielten allerdings die entsprechende Sequenz. Es wäre zu wünschen, daß Software, welche SAGE Projekte auswertet, auf derartige Tags automatisch aufmerksam macht.

Um diese Sonderfälle zu erfassen, wäre es notwendig, eine aufwendige interne Kontrolle durchzuführen - nämlich einen zweiten SAGE-Durchlauf mit einem anderen Verankerungsenzym beziehungsweise '*tagging*' Enzym.

Diese Zusammenfassung macht ersichtlich, daß SAGE als hoch sensibles Verfahren für diverse Störfaktoren anfällig ist, und es von eminenter Bedeutung ist, in zukünftigen Projekten einige Modifikationen umzusetzen beziehungsweise die Ursachen dieser Probleme weiter zu untersuchen.

Darüber hinaus existieren Unklarheiten im Vorgehen bei der Erstellung der endgültigen Taglisten und deren inhaltlicher Auswertung, was den Vergleich verschiedener Publikationen beziehungsweise von Transkriptomen verschiedener Arbeitsgruppen erschweren kann. Hier wäre eine genauere Darstellung in den Publikationen wünschenswert.

Mit der Umsetzung der in der vorliegenden Analyse genannten Aspekte und der ständigen Weiterentwicklung von SAGE (beispielsweise Lee et al. 2002, Saha et al. 2002) kann diese Methode ihr Potential noch weiter entfalten und dem Ziel, die (differentielle) Genexpression umfassend, digital und absolut an Stichproben zu messen, gerecht werden.

5 Statistische Evaluation und Reliabilität von SAGE

5.1 Ergebnisse

5.1.1 Ausgangsituation und Strategie

Um die Reliabilität von SAGE abschätzen zu können, wurde folgende Vorgehensweise gewählt.

Aus vier Mäusegroßhirnen von gesunden männlichen Tieren derselben Rasse, Alters,- und Gewichtsklasse war die Gesamt-RNS extrahiert und vereinigt worden. Diese Transkriptgrundpopulation war zweigeteilt und wie dargestellt parallel mittels SAGE untersucht worden, wobei pro Gruppe Zufallsstichproben von mehr als je 15000 Tags sequenziert wurden. Die nach Abzug der Artefakte (siehe S.70ff) verbleibenden 13548 (K1) beziehungsweise 13915 Tags (K2) waren einander anhand ihrer Sequenz zugeordnet worden.

Diese beiden Gruppen sollen nun statistisch auf Homogenität beziehungsweise ihren Zusammenhang (und dessen Ausmaß) untersucht werden. Zusätzlich zu dieser Ermittlung und Beurteilung der Reliabilität von SAGE sollen die Möglichkeiten der statistischen Auswertung von SAGE Projekten evaluiert werden. Hierzu wird die statistische Analyse "normaler" SAGE Experimente nachgeahmt. Die beiden Kontrollgruppen K1 und K2 werden dazu in zwei zu vergleichende Expressionsprofile umgedeutet. Das heißt, daß die Anzahl der Tagpaare, die auf dem α-Signifikanzniveau von 5% einen statistisch bedeutsamen Unterschied aufweisen, bestimmt werden soll. Um eventuelle Besonderheiten herauszuarbeiten, sollen zusätzlich bestimmte Untergruppen (Regulation mindestens zweifach, Mittelwert der Tagpaare mindestens fünf) untersucht werden. Die Ergebnisse dieser Berechnungen bilden die Grundlage für den statistischen Vergleich der Tests, die SAGE Daten paarweise testen (S. 127ff). Auf dieser Basis können Empfehlungen zur Verwendung der Tests abgegeben werden.

Die in der vorliegenden Arbeit entwickelte Sequenzfehlerkorrektur hatte eine zweite Version der beiden Profile K1 und K2 ($K1_{KORR}$ beziehungweise $K2_{KORR}$) ergeben. Diese soll mit dem nicht korrigierten, ursprünglichen Datensatz statistisch mit der Fragestellung verglichen werden, ob sich dadurch die Meßgenauigkeit verbessert (S. 130ff).

5.1.2 Vergleich der Gesamtverteilungen

In den nachstehenden Abschnitten werden die Verteilungen der beiden Expressionsprofile als Gesamtheit zur Bestimmung der Reliabilität anhand folgender Tests verglichen:

- Chi²- Test für *k* x 2-Felder-Tafeln (Simulationen),
- Kontingenzkoeffizient als Reliabilitätsmaß.

Daran schließen sich im darauffolgenden Kapitel die Einzelvergleiche der Tagpaare an.
Für sämtliche statistische Tests gelten folgende globale Hypothesen[14]:
H_o: Zwischen den beiden Stichproben (Expressionsprofile) K1 und K2 besteht bezüglich ihrer Merkmalsverteilung kein Unterschied.
H_1: Zwischen den beiden Stichproben (Expressionsprofile) K1 und K2 besteht bezüglich ihrer Merkmalsverteilung ein Unterschied.

5.1.2.1 Chi²-Test für k *x 2- Felder-Tafeln (Simulationen)*

Beschreibung des Tests

Zur Prüfung der Frage, ob die Gesamtverteilungen der beiden erstellten SAGE Profile K1 und K2 übereinstimmen (H_0), SAGE also als reliabel eingestuft werden kann, oder nicht (H_1), wurde folgende Vorgehensweise gewählt. Die globalen Hypothesen sollen mittels eines Testes zum Vergleich von Verteilungsfunktionen getestet werden. Für diskrete Verteilungen wie die vorliegenden kommt hierfür der Chi²-Test für *k* x 2 -Feldertafeln in Frage (Bortz 1993[4], S 159ff). Da als Voraussetzung für die gültige Anwendung dieses Testes gilt, daß die Erwartungswerte[15] größer oder gleich fünf sind (Bortz 1993[4], S. 159) und dies nur auf 1,8% der Zellen (Daten ohne Sequenzfehlerkorrektur) beziehungsweise 3,2% (Daten mit Korrektur) zutrifft, wurde eine Monte-Carlo Simulation durchgeführt.[16] Ausgangspunkt für diese Simulationen sind die beiden zu prüfenden Expressionsprofile. Es wird wiederholt eine Merkmalsverteilung erzeugt, die der H_0 entspricht, indem bei fixierten Spalten- und Zeilensummen die spezifischen Taghäufigkeiten n_1 und n_2 durch Zufallszahlen ersetzt werden. Anschließend wird der Chi²-Wert für diese simulierte Verteilung ermittelt. Mittels der Chi²-

---

[14] Wenn diese Hypothesen im Einzelfall durch spezifische Hypothesen untersetzt werden müssen, wird dies an gegebener Stelle dargestellt.

[15] Die Erwartungswerte *e* werden folgendermaßen berechnet: *e* = Zeilensumme x Spaltensumme/ Gesamttagzahl.

[16] Das entsprechende Programm in S-Plus 2000 wurde der Autorin freundlicherweise von E. H. Margulies (Department of Human Genetics, Universitity of Michigan Medical School, Ann Arbor, MI 48109, USA) zur Verfügung gestellt.

Werte der Iterationen kann eine spezifische Kennwertverteilung erstellt werden, welche die H_0-Verteilung darstellt. Anhand dieser Kennwertverteilung kann ermittelt werden, wie der empirische Chi²-Wert, das heißt der Kennwert der tatsächlich beobachteten Häufigkeitsverteilung, einzuschätzen ist. Läge er außerhalb der Verteilung, wäre H_1 anzunehmen.

Beschreibung der Berechnungen

Es wurden pro Datensatz 100 Zyklen durchgeführt. Es wurden für beide Datensätze (mit und ohne Sequenzfehlerkorrektur) die Gesamtverteilungen sowie die Verteilungen der Tagpaare mit einem Mittelwert $m \geq 5$ und ≥ 10 geprüft. Die Berechnung der empirischen Chi²-Werte für eine k x 2- Feldertafel erfolgte nach der Formel von Brandt-Snedecor (Sachs 1999[9], S. 585).

Aussagen zu den Hypothesen

Die H_0 kann nicht angenommen werden. In beiden Datensätzen (mit und ohne Sequenzfehlerkorrektur) liegen die Chi²-Werte der beobachteten Häufigkeitsverteilungen außerhalb der simulierten Verteilung. Dies gilt auch, wenn die Verteilungen der beiden Datensätze anhand der Kriteriums " $m \geq 5$ beziehungsweise 10" geprüft werden.
Nach Bortz (1990, S. 50) lautet die Entscheidungsregel zur globalen H_0, daß diese bereits abgelehnt werden muß, wenn ein *einziger* Test einen statistisch bedeutsamen Unterschied aufweist. Aus diesem Grund wird in der vorliegenden Arbeit auf die Durchführung der paarweisen Tests zum Reliabilitätsnachweis verzichtet.

Ergebnisse

Die ermittelten Verteilungen und die Chi²-Werte der beobachteten Verteilungen sind für die beiden Gesamtverteilungen (mit und ohne Korrektur) den Graphiken (Abb. 25 und 26) und den **Tabelle 14** und **15** für die Simulationen unter Ausschluß von Tagpaaren mit kleinen Mittelwerten zu entnehmen.

	Daten ohne Korrektur	**Daten mit Korrektur**
simulierte Chi²-Werte	217 - 360	334 - 467
beobachtete Chi²-Werte	368	628

Tabelle 14. Chi²-Werte der Verteilungen der Tagpaare mit einem Mittelwert von $m \geq 5$.

	Daten ohne Korrektur	Daten mit Korrektur
simulierte Chi²-Werte	66 - 131	97 - 180
beobachtete Chi²-Werte	170	204

Tabelle 15. Chi²-Werte der Verteilungen der Tagpaare mit einem Mittelwert von $m \geq 10$.

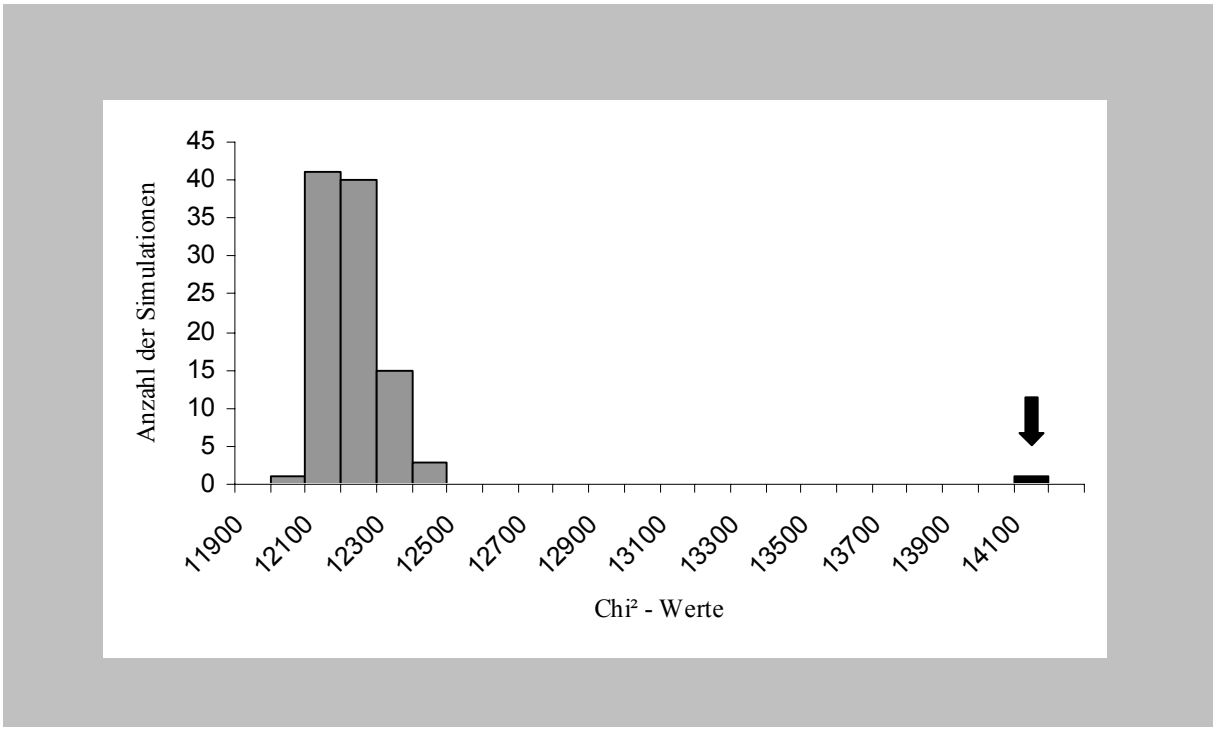

Abb. 25: Monte-Carlo Simulationen der gesamten Daten mit Korrektur. Das Histogramm zeigt die simulierte Verteilung der Chi² Werte unter der Annahme der Nullhypothese (graue Balken). Der Pfeil weist auf die Lage des beobachteten Chi² Wertes hin (schwarzer Balken).

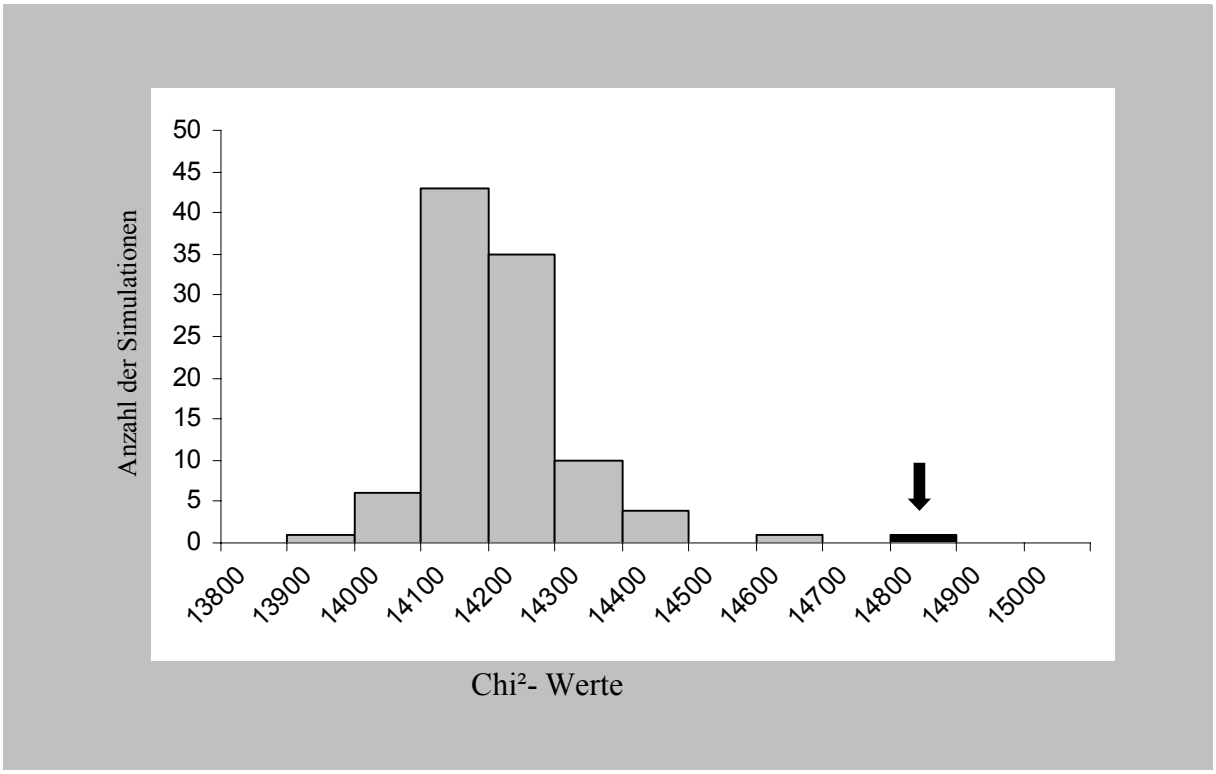

Abb. 26: Monte-Carlo Simulationen der gesamten Daten ohne Korrektur. Das Histogramm zeigt die simulierte Verteilung der Chi² Werte unter der Annahme der Nullhypothese (graue Balken). Der Pfeil weist auf die Lage des beobachteten Chi² Wertes hin (schwarzer Balken).

5.1.2.2 Kontingenzkoeffizient

Beschreibung

Nach Bortz (1990, S. 60) kann das Ausmaß der Reliabilität für nominalskalierte Merkmale anhand eines Kontingenzkoeffizienten beschrieben werden. Dieser Kennwert gibt das Maß der Enge des Zusammenhangs zwischen den Merkmalen der entsprechenden Kontingenztafel wieder und basiert auf χ^2-Techniken. Als Voraussetzung gilt, daß die Existenz eines statistisch signifikanten Zusammenhangs gesichert ist. Die Hypothesen, die hierfür die Grundlage bieten, beziehen sich nicht auf die Gleichartigkeit der Merkmalsverteilungen der beiden Kontrollgruppen (siehe S. 113), sondern auf die Prüfung ihrer stochastischen (Un-) Abhängigkeit[17]. Es sind also folgende Hypothesen zu überprüfen:

[17] Für die statistischen Berechnungen ist dieser Unterschied unerheblich, es liegen ihm jedoch unterschiedliche Zufallsmodelle zugrunde. Während Zusammenhangshypothesen die Realisierung einer bivariaten Zufallsvariablen an einer Stichprobe untersuchen, wird anhand von Unterschiedshypothesen die Realisierung einer univariaten Zufallsvariablen an zwei Stichproben untersucht (Bortz 1990, S. 103). In diesem Abschnitt wird also im Sinne einer Meßwiederholung das ursprüngliche Boten-RNS-Pool als eine Stichprobe aufgefaßt.

H_0:. Die beide Kontrollgruppen sind voneinander stochastisch unabhängig.

H_1: Die beide Kontrollgruppen sind voneinander stochastisch abhängig.

Die Kontingenztafel, die dem Hypothesen-Test und dem Koeffizienten zugrunde liegt, gestaltet sich folgendermaßen:

K2 / **K1**	**1**	**2**	**3**	**etc.**
1	x_1	x_2	x_3	x…
2	x_4	x_5	x_6	x…
3	x_7	x_8	x_9	x…
etc.	x…	x...	x...	x_n

Tabelle 16. Prototypische Kontingenztafel. Diese Tafel dient der Gegenüberstellung der beiden Kontrollgruppen. Bei den Benennungen "1, 2, 3, etc. " handelt es sich um die Taghäufigkeiten, die Werte x_1 bis x_n geben die Anzahl der Tagpaare wieder, welche die jeweilige Konstellation aufweisen.

Als Assoziationsmaß wurde Cramérs V gewählt. Dieser Kennwert ermittelt sich nach folgender Formel:

$$\mathbf{V} = \sqrt{\frac{\chi^2}{N(k-1)}} \qquad \text{(Gleichung 2)}$$

Hierbei ist k = min (Spaltenanzahl, Zeilenanzahl) und N die Summe der Beobachtungen (Tagpaare). Des weiteren wurde getestet, ob sich V von dem Wert 0 statistisch signifikant unterscheidet.

Beschreibung der Berechnungen

Es wurden jeweils für beide Datensätze (mit und ohne Sequenzfehlerkorrektur) beide Kontrollgruppen einander gegenübergestellt. Dabei wurde folgendes Vorgehen gewählt: Zuerst wurden die Kontingenztafeln per Chi²-Test auf Gültigkeit der Hypothesen beziehungsweise Existenz eines Zusammenhangs (H_1) überprüft. Wenn dieser als statistisch gesichert angesehen werden konnte, wurde Cramérs V berechnet und auf statistische Signifikanz überprüft. Dieses Vorgehen wurde für sämtliche Daten, deren Mittelwert $m \geq 5$ beziehungsweise $m < 5$ war, wiederholt. Die Erstellung der Kontingenztabellen und alle Berechnungen erfolgten anhand statistischer Software (SPSS).

Aussagen zu den Hypothesen

Die H_1, daß K1 und K2 stochastisch abhängig sind, ist auf dem 1% Niveau für alle untersuchten Konstellationen abgesichert. Zwischen beiden Gruppen besteht also ein

statistisch bedeutsamer Zusammenhang. Die zugehörigen Werte sind **Tabelle 17** zu entnehmen.

	Chi²-Wert	Freiheitsgrade	*p*
ohne: alle	249644,5	1587	$\rightarrow 0$
$m \geq 5$	4402,444	1404	$\rightarrow 0$
$m < 5$	11133,57	64	$\rightarrow 0$
mit: alle	183006,8	1591	$\rightarrow 0$
$m \geq 5$	5806,633	1548	$\rightarrow 0$
$m < 5$	8683,832	81	$\rightarrow 0$

Tabelle 17. Chi²-Werte, Freiheitsgrade und Quantil der Kontingenztafeln. Die Bezeichnungen "ohne" beziehungsweise "mit" kennzeichnen die Datensätze mit beziehungsweise ohne Sequenzfehlerkorrektur, "alle" bezieht sich auf jeweils den gesamten Datensatz, "$m < / \geq 5$" auf die selektierten Datensätze, deren Mittelwert größer beziehungsweise kleiner 5 ist.

Ergebnisse

Tabelle 18 sind die berechneten Werte von Cramérs V für die Gegenüberstellung von K1 und K2 ohne Selektion sowie getrennt für Tagpaare, deren Mittelwert größer oder gleich beziehungsweise kleiner fünf ist, zu entnehmen. Sämtliche Koeffizienten sind statistisch signifikant von 0 verschieden ($p \rightarrow 0$). Da die Bewertung dieser Ergebnisse sehr komplex ist, wird auf eine Darstellung an dieser Stelle verzichtet und auf die Einschätzung der Ergebnisse in bezug auf die Meßgenauigkeit von SAGE im Rahmen der Diskussion verwiesen (siehe S. 162f).

Auswahl	Daten ohne Sequenzfehlerkorrektur	Daten mit Sequenzfehlerkorrektur
alle	0,681	0,637
$m \geq 5$	0,693	0,642
$m < 5$	0,316	0,286

Tabelle 18. Cramérs V. Die Bezeichnung "alle" bezieht sich auf jeweils den gesamten Datensatz, "$m < / \geq 5$" auf die selektierten Datensätze, deren Mittelwert größer beziehungsweise kleiner 5 ist.

5.1.3 Paarweise Vergleiche

Nach Bortz (1990, S. 50) lautet die Entscheidungsregel zu globalen Hypothesen (H_0), daß diese bereits abgelehnt werden müssen, wenn ein *einziger* Test einen statistisch bedeutsamen Unterschied aufweist. Da der den paarweisen Tests vorangehende Gesamtvergleich der beiden Verteilungen bereits zur Ablehnung der H_0 führte und so die Aussage zur globalen H_0 bereits entschieden ist, wird in der vorliegenden Arbeit auf die Durchführung der paarweisen Tests zum Reliabilitätsnachweis verzichtet.

In den üblichen SAGE Experimenten[18] dagegen stehen die einzelnen Hypothesen zu jedem Tagpaar im Mittelpunkt des Forschungsinteresses zur differentiellen Regulation der Genexpression. Um die möglichen Ergebnisse eines solchen üblichen SAGE Experimentes zu simulieren und so Empfehlungen für den statistischen Entscheidungsprozeß üblicher SAGE Experimente abgeben zu können, werden beispielhaft die auf dem 5% Niveau statistisch als unterschiedlich zu betrachtenden Tagpaare ermittelt und unter den Aspekten "praktisch relevanter Unterschied (Regulation)" und "minimale Taghäufigkeiten" evaluiert.

In diesen Fällen gelten als paarweise Hypothesen: H_o: Zwischen den beiden Tags eines Paares besteht bezüglich ihrer Häufigkeiten kein Unterschied. H_1: Zwischen den beiden Tags eines Paares besteht bezüglich ihrer Häufigkeiten ein Unterschied. Sämtliche Berechnungen werden für beide Datensätze (mit und ohne Sequenzfehlerkorrektur) durchgeführt.

Es werden die folgenden Tests angewandt:

- Test nach Madden et al. (1997),
- Test nach Audic und Claverie (1997),
- Chi²- Test für Vier-Felder-Tafeln,
- SAGEstat (Kal et al. 1999) beziehungsweise modifizierter Z-Test.

5.1.3.1 Test nach Madden et al. (1997)

Beschreibung

Ebenso wie Audic und Claverie (1997) (siehe unten) wählen Madden et al. (1997) einen auf der Poissonverteilung (Kal et al. 1999) beruhenden Ansatz. Die auf den spezifischen Häufigkeitswerten des interessierenden Tagpaares basierende Formel für statistisch signifikante Unterschiede lautet:

---

[18] Der Begriff "übliche SAGE Experimente" meint hier und im weiteren SAGE Projekte, die Expressionsprofile von zwei verschiedenen Geweben, Zuständen etc. miteinander vergleichen, um Unterschiede in der Genexpression herauszuarbeiten.

$$\mathbf{0 < (x - k * x0{,}5) - (y + k * y0{,}5),} \qquad \text{(Gleichung 3)}$$

Bei x und y handelt es sich um die spezifischen Tagmengen eines Paares und bei *k* um einen Faktor, der das gewählte Signifikanzniveau integriert und damit den Konfidenzgrad darstellt. Der Faktor *k* entspricht dem erwünschten Signifikanzniveau mit 1,96 bei $p = 0{,}05$[19]. X muß größer als y gewählt werden. Die Entscheidungsregel lautet: Die H_o, daß zwischen x und y keine Differenz besteht, kann abgelehnt werden, wenn Gleichung 2 positive Werte ergibt[20]. In der vorliegenden Arbeit wurden die Berechnungen zur Signifikanz nach Madden et al. (1997) nach der von Ruijter (1999) umgeformten Formel durchgeführt:

$$\mathbf{a = |x\text{-}y| / (x0{,}5 + y0{,}5)} \qquad \text{(Gleichung 4).}$$

Dem Kennwert *a* wurde das entsprechende Quantil der Standardnormalverteilung zugeordnet und mit $\alpha/2$ verglichen.

Simulation der statistischen Entscheidungen üblicher SAGE Experimente

Tabelle 19 ist zu entnehmen, wie viele Tagpaare einen statistisch bedeutsamen Unterschied aufgewiesen hätten.

Signifikanz-niveau	Daten ohne Sequenzfehlerkorrektur				Daten mit Sequenzfehlerkorrektur			
	Gesamt	≥ 2fach	$m \geq 5$	beides	Gesamt	≥ 2 fach	$m \geq 5$	beides
5%	96 (0,7%)	94 (6,6%)	2 (0,8%)	0 (0%)	272 (2,2%)	270 (17,9%)	17 (4,3%)	15 (14,6%)
Anzahl der Paare insgesamt	14159	1432	255	64	12182	1511	391	103

Tabelle 19. Ergebnisse des Tests nach Madden et al. (1997). Es sind für beide Datensätze die Anzahl der Tagpaare (in Klammern: prozentualer Anteil an der jeweiligen Bezugsgruppe) aufgeführt, für die unter dem Signifikanzniveau α die H_0 verworfen werden muß. Die Spalte "Gesamt" enthält die Daten der Gesamtprofile, "≥2fach" enthält nur die Daten der Tagpaare, die mindestens einen Häufigkeitsunterschied um den Faktor zwei aufweisen, "$m \geq 5$", die mindestens einen Mittelwert von fünf besitzen, und "beides" enthält die Daten der Paare, die beide Kriterien erfüllen.

[19] Dieser Faktor beruht somit auf der Z-Verteilung (Standardnormalverteilung).

[20] Dies kann in folgende Form (Ruijter 1999) umgewandelt werden: H_o wird abgelehnt, wenn $(x - y) / (x^{0,5} + y^{0,5}) > Z_{\alpha/2}$ ist.

5.1.3.2 Test nach Audic und Claverie (1997)

Beschreibung des statistischen Ansatzes

Die Veröffentlichung von Audic und Claverie (1997) zur Berechnung der Signifikanz digitaler Genexpressionsdaten beruht auf einem klassischen statistischen Ansatz, der dem Sammelprozeß von SAGE die Poissonverteilung zugrunde legt[21]. Sie entwickeln folgende Gleichung zur Berechnung der Wahrscheinlichkeit, zwei Vorfälle *x* und *y* (Häufigkeiten eines bestimmten Tags in zwei Expressionsprofilen) zu beobachten, die auf dem gleichen seltenen Ereignis (Boten-RNS - Niveau) beruhen[22], wobei die Anzahl der pro Profil insgesamt sequenzierten Tags (N_1 und N_2) unterschiedlich sein darf und mit in die Formel eingeht:

$$p(y|x) = \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!\,y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}} \qquad \text{(Gleichung 5)}$$

$P(y|x)$ gibt die bedingte Wahrscheinlichkeit an, mit der bei Gültigkeit von H_o[23] erwartet wird, die Häufigkeit *y* eines bestimmten Tags zu beobachten, wenn in einem anderen Profil dieses Tag bereits *x*- mal aufgetreten ist. Unter Verwendung von Gleichung 5 berechnen die Autoren durch Aufsummierung[24] für die gegebenen *x* Werte Konfidenzintervalle $[y_{min}, y_{max}]$, innerhalb derer der Wert von *y* mit einer Wahrscheinlichkeit p = 1 - α (beispielsweise 95%) liegen sollte. Diese Intervalle tragen der Tatsache Rechnung, daß es sich bei SAGE um einen zufälligen Sammelprozeß handelt, der auch dann einer gewissen Fluktuation unterworfen ist,

---

[21] Diese gibt die Verteilung seltener Ereignisse wieder. Das heißt, daß die Anzahl aller Ereignisse n (Transkripte einer RNS-Population) sehr groß ist und die Wahrscheinlichkeit p des untersuchten Alternativereignisses (konkretes Transkript) sehr klein. Daraus folgt, daß die exakte binomiale Wahrscheinlichkeitsfunktion durch die Poisson-Verteilung approximiert werden kann. Für $N \to \infty$ und $p \to 0$ geht die Binomial- in die Poissonverteilung über (Bortz 1993, S. 70f). Audic und Claverie (1997) gehen davon aus, daß die einzelnen Transkripttypen jeweils nicht mehr als 5% einer RNS-Population ausmachen und somit als seltene Ereignisse betrachtet werden können.

[22] Audic und Claverie gehen davon aus, daß die apriori Wahrscheinlichkeit für alle Ereignisse im Bereich von Null bis Unendlich gleich ist.

[23] Hier paarspezifische H_0: Die beobachteten Taghäufigkeiten *x* und *y* sind gleich. H_1: Sie unterscheiden sich.

[24] Die Grenzen der Konfidenzintervalle lauten:

$$C\ (y \le y_{min}|x) = \sum_{y=0}^{y \le y_{min}} p(y \mid x) \text{ und } D\ (y \ge y_{max} \mid x) = \sum_{y=y_{max}}^{\infty} p(y \mid x)$$

wenn den Tags zweier Expressionsprofile ein unreguliertes Gen zugrunde liegt. Liegt ein *y*-Wert innerhalb des entsprechenden Konfidenzintervalls, so spricht dies für die Gültigkeit von H_0. Befindet er sich außerhalb davon, muß die Alternativhypothese auf dem entsprechenden Signifikanzniveau angenommen werden. In diesem Fall scheinen stochastisch sehr unwahrscheinliche Schwankungen zwischen den beiden Profilen vorzuliegen.

Beschreibung der Berechnungen

Es wurde folgendes Vorgehen gewählt: Es wurden Tafeln erstellt, in welchen zu sämtlichen gegebenen *x*- Werten eines Datensatzes (entspricht K1) die Summenwahrscheinlichkeiten der *y*-Werte von 0 bis mindestens zu dem beobachteten *y*-Wert (K2) errechnet wurden. So konnte die summierte bedingte Wahrscheinlichkeit für ein beobachtetes Tagpaar abgelesen und mit $\alpha/2$ verglichen werden[25].

Simulation der statistischen Entscheidungen üblicher SAGE Experimente

Tabelle 20 kann entnommen werden, wie viele Tagpaare einen statistisch signifikanten Unterschied gezeigt hätten. Diese Daten wurden nicht nur für die Gesamtprofile mit und ohne Korrektur ermittelt, sondern auch für bereits genannten Untergruppen der beiden Datensätze. Einmal wurden alle Tagpaare ausgewählt, die mindestens zweifach verschieden waren, dann diejenigen, die mindestens einen Mittelwert von fünf aufwiesen, und zuletzt diejenigen, die beiden Kriterien entsprachen.

[25] Da hier zwei einseitige kumulative Wahrscheinlichkeiten ermittelt werden, muß als Signifikanzniveau $\alpha/2$ zum Vergleich herangezogen werden (Man et al. 2000).

Signifikanz-niveau	Daten ohne Sequenzfehlerkorrektur				Daten mit Sequenzfehlerkorrektur			
	Gesamt	≥ 2fach	$m \geq 5$	beides	Gesamt	≥ 2 fach	$m \geq 5$	beides
5%	67 (0,5%)	59 (4,1%)	24 (9,4%)	16 (25,0%)	243 (2,0%)	236 (15,6%)	53 (13,6%)	46 (44.7%)
Anzahl der Paare insgesamt	14159	1432	255	64	12182	1511	391	103

Tabelle 20. Ergebnisse des Tests nach Audic und Claverie. Es sind für beide Datensätze die Anzahl der Tagpaare (in Klammern: prozentualer Anteil an der jeweiligen Bezugsgruppe) aufgeführt, für die unter dem Signifikanzniveau α die H_0 verworfen werden muß. Die Spalte "Gesamt" enthält die Daten der Gesamtprofile, "≥2fach" enthält die Daten der Tagpaare, die mindestens einen Häufigkeitsunterschied um den Faktor zwei aufweisen, "$m \geq 5$", die mindestens einen Mittelwert von fünf besitzen, und "beides" enthält die Daten der Paare, die beide Kriterien erfüllen.

5.1.3.3 Vier-Felder-Chi²-Test

Beschreibung

Methoden, die auf χ^2-Prüfstatistiken basieren, dienen der Analyse von Häufigkeitsunterschieden im Auftreten bestimmter Merkmale (Bortz, 1993[4], S. 145). Sie sind also für SAGE geeignet, ohne ein spezifisches Verfahren darzustellen, das speziell für SAGE entwickelt worden ist.

Ausgehend von den Darstellung in Man et al. (2000) und der einer UniGene-Internetseite (www.ncbi.nlm.nih.gov/UniGene/fisher.shtml, 1.10.2002) wurden für den Vergleich einzelner Tagpaare jeweils folgende Vier-Felder-Tafeln erstellt:

	K1	K2	Summen
Tagpaar	a	c	N_{TAG}
Rest	b	d	N_{REST}
Summen	N_1	N_2	N_{GES}

Tabelle 21. Vier-Felder-Tafel für den Chi²-Test. Bei *a* und c handelt es sich um die beobachteten Häufigkeiten eines bestimmten Tagpaares, *b* und *d* entstehen jeweils durch Subtraktion von *a* beziehungsweise *c* von N_1 beziehungsweise N_2.

Der statistische Kennwert χ^2 wurde anhand folgender Formel berechnet (Sachs 1999[9], S. 451):

$$\chi^2 = \frac{N_{GES}(ad - bc)^2}{N_1 N_2 N_{TAG} N_{REST}} \qquad \text{(Gleichung 6)}$$

Daran schloß sich die Ermittlung des entsprechenden Quantils der χ^2-Verteilung für einen Freiheitsgrad an.

Simulation der statistischen Entscheidungen üblicher SAGE Experimente

Der Tabelle 22 kann entnommen werden, wie viele Tagpaare einen statistisch signifikanten Unterschied gezeigt hätten. Es wurde dabei folgendes Vorgehen gewählt: Diese Daten wurden nicht nur für die Gesamtprofile mit und ohne Korrektur ermittelt, sondern auch für die drei bereits genannten Untergruppen der beiden Datensätze. Einmal wurden alle Tagpaare ausgewählt, die mindestens zweifach verschieden waren, dann diejenigen, die mindestens einen Mittelwert von fünf aufwiesen, und zuletzt diejenigen, die beiden Kriterien entsprachen.

Signifikanz-niveau	Daten ohne Sequenzfehlerkorrektur				Daten mit Sequenzfehlerkorrektur			
	Gesamt	≥ 2fach	$m \geq 5$	beides	Gesamt	≥ 2 fach	$m \geq 5$	beides
5%	129 (0,9%)	121 (8,5%)	26 (10,2%)	18 (28,1%)	333 (2,7%)	326 (21,6%)	54 (13,8%)	47 (45,6%)
Anzahl der Paare insgesamt	14159	1432	255	64	12182	1511	391	103

Tabelle 22. Ergebnisse des Vier-Felder-Chi²-Tests. Es sind für beide Datensätze die Anzahl der Tagpaare (in Klammern: prozentualer Anteil an der jeweiligen Bezugsgruppe) aufgeführt, für die unter dem Signifikanzniveau α die H_0 verworfen werden muß. Die Spalte "Gesamt" enthält die Daten der Gesamtprofile, "≥2fach" enthält die Daten der Tagpaare, die mindestens einen Häufigkeitsunterschied vom Faktor zwei aufweisen, "$m \geq 5$"**,** die mindestens einen Mittelwert von fünf besitzen, und "beides" enthält die Daten der Paare, die beide Kriterien erfüllen.

5.1.3.4 Z-Test

Beschreibung

Der Z-Test zur Prüfung der Gleichheit zweier Proportionen wurde von Kal et al. (1999) vorgestellt. Dieser Ansatz betrachtet die Anzahl der Kopien einer bestimmten Boten-RNS in einer Zelle als Bruchteil aller Boten-RNS Moleküle in dieser Zelle. Dieser spezifische Anteil sollte in der SAGE Bibliothek mit dem Verhältnis dieses bestimmten Transkriptes zu allen

sequenzierten Tags übereinstimmen. Für die große Anzahl an sequenzierten Tags geht diese Binomialverteilung in eine Normalverteilung über (Kal et al. 1999). Aus der Differenz der beiden Proportionen p_1 und p_2 eines Tagpaars ($\frac{n_1}{N_1} - \frac{n_2}{N_2}$) und ihres Standardfehlers wird folgende Teststatistik entwickelt:

$$\mathbf{z} = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)(\frac{1}{N_1} + \frac{1}{N_2})}} \qquad \text{(Gleichung 7),}$$

wobei $p_0 = (n_1 + n_2) / (N_1 + N_2)$[26] ist und den Schätzwert der Proportionen unter der Bedingung, daß die H_0 wahr ist, darstellt. H_0 ist zu verwerfen, wenn entweder $z > z_{\alpha/2}$ oder wenn $z < - z_{\alpha/2}$ ist[27]. Nach Sachs (1999[9], S. 441) sollte dieser Test nur angewandt werden, wenn folgende Konstellationen zutreffen: a) $N_1 \geq 50$ und $N_2 \geq 50$, b) $N_1 p_1 > 5$ und $N_2 p_2 > 5$, c) $N_1 (1 - p_1) > 5$ und $N_2 (1 - p_2) > 5$. Dies ist bei SAGE Experimenten oft nicht gegeben.
Z läßt sich deswegen alternativ nach Sachs (1999[9]) exakter und vor allem unter weniger strengen Voraussetzungen anhand einer auf der Winkeltransformation beruhenden Approximation berechnen:

$$\mathbf{z} = \frac{\left|\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2}\right|}{28{,}648\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \qquad \text{(Gleichung 8)}$$

Hierfür gelten folgende Bedingungen: a) $N_1 \geq 25$ und $N_2 \geq 25$, b) $N_1 p_1 > 1$ und $N_2 p_2 > 1$, c) $N_1 (1 - p_1) > 1$ und $N_2 (1 - p_2) > 1$ (Sachs 1999[9]).

Simulation der statistischen Entscheidungen üblicher SAGE Experimente

Da die Ergebnisse des *nicht modifizierte* Test exakt mit denjenigen des Vier-Felder-Chi²-Testes übereinstimmen, wurde auf eine gesonderte Darstellung verzichtet.
In Tabelle 23 sind die Resultate des nach Sachs (1999[9]) modifizierten Z-Testes dargestellt.

---

[26] Man et al. (2000) geben die Formel für p_0 leicht verändert wieder, was zu einer minimalen Veränderung der Ergebnisse führt.

[27] Hierbei handelt es sich um die entsprechenden Quantile der Standardnormalverteilung.

Signifikanz-niveau	Daten ohne Sequenzfehlerkorrektur				Daten mit Sequenzfehlerkorrektur			
	Gesamt	≥ 2fach	$m \geq 5$	beides	Gesamt	≥ 2 fach	$m \geq 5$	beides
5%	1351 (9,5%)	842 (58,8%)	37 (14,5%)	28 (43,8%)	1548 (12,7%)	998 (66,0%)	70 (17,9%)	63 (61,2%)
Anzahl der Paare insgesamt	14159	1432	255	64	12182	1511	391	103

Tabelle 23. Ergebnisse des modifizierten Z-Testes. Es sind für beide Datensätze die Anzahl der Tagpaare (in Klammern: prozentualer Anteil an der jeweiligen Bezugsgruppe) aufgeführt, für die unter dem Signifikanzniveau α die H_0 verworfen werden muß. Die Spalte "Gesamt" enthält die Daten der Gesamtprofile, "≥2fach" enthält die Daten der Tagpaare, die mindestens einen Häufigkeitsunterschied vom Faktor zwei aufweisen, "$m \geq 5$", die mindestens einen Mittelwert von fünf besitzen, und "beides" enthält die Daten der Paare, die beide Kriterien erfüllen.

5.1.4 Zusammenfassung der Ergebnisse der Simulationen üblicher Experimente

Im Folgenden sollen diese Ergebnisse der vier paarweisen Tests zueinander in Bezug gestellt werden.

Im Rahmen der Simulationen üblicher SAGE Experimente war herausgearbeitet worden, welchen Effekt es auf die Ergebnisse hat, wenn die Expressionsprofile nach den Kriterien "minimale Verschiedenheit" (Regulation) und "minimale Taghäufigkeit" selektiert werden. Bei allen vier Tests ist eine leichte (Regulation) bzw. starke ($m \geq 5$) Abnahme der Anzahl der statistisch signifikant verschiedenen Tagpaare zu beobachten. Die stärkste Reduktion findet bei der Anwendung beider Kriterien statt. Im Fall des modifizierten Z-Tests und des Tests nach Madden et al. (1997) ist die Reduktion der Anzahl der statistisch signifikant verschiedenen Tagpaare besonders ausgeprägt, wenn die Daten anhand des Kriteriums $m \geq 5$ ausgewählt werden; wohingegen die Selektion von Tagpaaren, die mindestens eine zweifachen Unterschied aufweisen, nur eine minimale Reduktion bewirkt. Um zu ermitteln, ob es solche und weitere testspezifischen Unterschiede von statistischer Relevanz gibt, werden die Resultate der verschiedenen Tests im nächsten Abschnitt rechnerisch verglichen.

Auffällig ist des weiteren, daß sich die Resultate der Daten ohne Sequenzfehlerkorrektur von denjenigen mit Korrektur teilweise stark unterscheiden. Diese Unterschiede sollen weiter unter statistisch evaluiert werden (siehe S. 130ff).

5.1.5 Statistischer Vergleich der angewandten paarweisen Tests

Es soll geprüft werden, ob die Ergebnisse der Simulation üblicher SAGE Experimente anhand verschiedener Test statistisch signifikante Unterschiede aufweisen. Dazu wurde folgendes Vorgehen gewählt. Um eventuelle Unterschiede zwischen den einzelnen Tests herauszuarbeiten, wurden diese einzeln miteinander verglichen. Dabei wurden folgende Hypothesen geprüft:

H_0: Der Anteil an Tagpaaren, die auf dem geprüften Niveau einen statistisch signifikanten Unterschied aufweisen, ist in beiden Tests gleich.

H_1: Der Anteil an Tagpaaren, die auf dem geprüften Niveau einen statistisch signifikanten Unterschied aufweisen, ist in beiden Tests unterschiedlich.

Die beiden Datensätze (mit und ohne Sequenzfehlerkorrektur) wurden getrennt betrachtet.

Es wurde auf eine Fehleradjustierung verzichtet. Da grundsätzlich eine komplexe statistische Situation im Sinne von multiplen Testen vorliegt, ist damit das Vorgehen explorativ zu verstehen. Das heißt, daß die Ergebnisse nicht als allgemein gültig und konfirmativ aufgefaßt werden dürfen.

Beschreibung des statistischen Vergleichs

Zur Überprüfung der beiden Hypothesen wurde mittels 4-Felder-Chi²-Test die Anzahl der Tagpaare, die einen statistisch bedeutsamen Unterschied aufweisen, verglichen. Die entsprechende Tafel gestaltet sich folgendermaßen:

	verschieden	nicht verschieden	Summen
Test 1	a	c	N_{Paare}
Test 2	b	d	N_{Paare}
Summen	$N_{verschieden}$	$N_{nicht\ verschieden}$	N

Tabelle 24. Vier-Felder-Tafel zum Testvergleich. Bei *a* und *b* handelt es sich um die Anzahl der ermittelten Paare, die einen statistisch signifikanten Unterschied aufweisen, bei *c* und *d* um die Anzahl der entsprechenden nicht statistisch signifikant verschiedenen Paare. N_{Paare} hat in dem Datensatz ohne Korrektur immer einen Wert von 14159, in dem Datensatz mit Korrektur von 12182.

Der Test überprüft, ob die Besetzung der Felder homogen ist oder statistisch signifikant verschieden. Es wurde auf dem 5% Niveau getestet. Wenn sich eine statistisch bedeutsame inhomogene Besetzung der Felder ergab, wurden korrigierte standardisierte Residuen

(Abweichungen der beobachteten von den erwarteten Werten[28]) berechnet, um erkennen zu können, welcher Test wie zur Signifikanz beiträgt.

Aussagen zu den Hypothesen

Die H_1 kann auf dem 5% Niveau für folgende Testergebnisse in beiden Datensätzen als gesichert gelten: sämtliche Kombinationen, die den modifizierten Z-Test enthalten, sowie der Vergleich der Ergebnisse des Chi²-Tests mit denjenigen des Tests von Audic und Claverie. Die H_1 kann für folgende Testergebnisse in beiden Datensätzen nicht angenommen werden: sämtliche Konstellationen außer den oben genannten.

Ergebnisse

Der folgenden Tabelle (**Tabelle 25**) sind die Testkonstellationen und ihre statistische Bewertung zu entnehmen. Da sich die Ergebnisse für die beiden Datensätze (mit und ohne Sequenzfehlerkorrektur) nicht unterscheiden, werden die Resultate des Testsvergleichs nur einmal dargestellt.

Simulation 5%	**Audic und Claverie**	**Madden et al.**	**4-Felder-Chi²**
Madden	gleich	-	-
4-Felder-Chi²	verschieden (4,4 / 3,8)	gleich	-
modifizierter Z	verschieden (35 / 32)	verschieden (33,9 / 31,1)	verschieden (32,6 / 29,2)

Tabelle 25. Vergleich der Ergebnisse der paarweisen Tests. "Gleich" weist auf die Annahme von H_0 hin, während "verschieden" die Bestätigung von H_1 wiedergibt. In den Klammern sind die korrigierten standardisierten Residuen (*z*-Werte) derjenigen Zellen angegeben, welche die Anzahl der statistisch signifikant verschiedenen Tagpaare des in dieser Tabelle links aufgeführten Testes enthalten (erste Zahl: Werte der Daten ohne Sequenzfehlerkorrektur, zweite: Datensatz mit Korrektur).

Der modifizierte Z-Test unterscheidet sich auf allen drei simulierten Signifikanzniveaus von allen anderen Tests statistisch signifikant (immer *p* gegen 0). Die Analyse der entsprechenden korrigierten standardisierten Residuen zeigt, daß im Vergleich zu den anderen drei Testarten bei diesem Test überzufällig viele Tagpaare statistisch signifikant werden. Die entsprechenden *z*-Werte sind **Tabelle 25** zu entnehmen (z_{KRIT} = 1,645)[29]. Der Vergleich des

[28] Es handelt sich um approximativ normalverteilte standardisierte Residuen, die aus dem Quotienten aus der Differenz des beobachteten und dem erwarteten Wert und dem geschätzten Standardfehler gebildet werden. Zur Signifikanzprüfung wurde der dem α-Wert zugehörige kritische *z*-Wert berechnet (im Folgenden z_{KRIT} genannt).

[29] Der nach Bonferoni adjustierte α-Wert würde folgendermaßen lauten: α_{KORR} = 0,05 / 24 = 0,0021. Daraus

Chi²-Tests mit dem Test nach Audic und Claverie zeigt einen statistisch bedeutsamer Unterschied mit einer Überrepräsentation der mittels Chi²-Test als verschieden ermittelten Tagpaare. Im folgenden sollen diese Konstellationen näher beleuchtet werden.

Spezieller Vergleich des Chi²-Tests mit dem Test nach Audic und Claverie

Man et al. (2000) berichten, daß der Vier-Felder-Chi²-Test im Bereich kleiner Taghäufigkeiten ($n < 15$) eine höhere Teststärke besitzt als der Test nach Audic und Claverie. Das bedeutet, daß der Chi²-Test eine höhere Wahrscheinlichkeit aufweist, tatsächlich vorhandene Unterschiede auch aufzudecken. Um zu prüfen, ob der beobachtete Unterschied zwischen den beiden Tests auf dieser Beobachtung beruhen könnte, wurde die Anzahl der auf dem 5% Simulationsniveau statistisch signifikanten Tagpaare, deren Mittelwert $m \geq$ beziehungsweise $m < 15$ für jeden Test ermittelt und anhand des Z-Test auf dem 5% Niveau verglichen.[30] Dies ergab: Die beiden Test entscheiden in beiden Datensätzen dann identisch, wenn nur die Anzahl der statistisch signifikanten Tagpaare betrachtet wird, die einen Mittelwert m von ≥ 15 haben. Die H_1 kann hier also nicht angenommen werden. Wird dagegen die Anzahl der statistisch signifikanten Tagpaare betrachtet, die einen Mittelwert m von < 15 haben, kann die H_1 bestätigt werden ($p \rightarrow 0$). Die entsprechenden Verteilungen der Anzahl der Tagpaare sind **Tabelle 26** zu entnehmen.

	Daten ohne Sequenzfehlerkorrektur		Daten mit Sequenzfehlerkorrektur	
	Audic	Chi	Audic	Chi
$m \geq 15$	13	13	14	14
$m < 15$	54	116	229	319

Tabelle 26. Vergleich des Tests nach Audic und Claverie mit dem Chi²-Test. Es ist die Anzahl der auf dem simulierten 5% Niveau signifikanten Tagpaare in Abhängigkeit von deren Häufigkeit aufgeführt. "*m*" bezeichnet den Mittelwert.

Die beiden Tests weisen also ein deutlich verschiedenes Entscheidungsverhalten in Abhängigkeit der geprüften Taghäufigkeiten auf. Die höhere Teststärke der Chi²-Testes im Bereich kleiner Taghäufigkeiten zeigt sich deutlich.

---

ergäbe sich $z_{KRIT} = 2{,}87$. An der Aussage würde sich dadurch nichts ändern.

[30] Die Hypothesen sind mit denjenigen zum allgemeinen Testvergleich identisch .

Vergleich der Ergebnisse des modifizierten Z-Test mit denjenigen der anderen Test

Es sollte geprüft werden, ob sich die Ergebnisse des modifizierten Z-Tests auch dann von denjenigen der anderen drei Tests unterscheiden, wenn die Anzahl der statistisch signifikanten Paare verglichen wird, die sich ergibt, wenn nur Paare betrachtet werden, die einen Mittelwert von $m \geq 5$ haben. Dieser Vergleich erfolgte ebenfalls analog zu dem auf S. 127f vorgestellten. Die Ergebnisse sind Tabelle 27 zu entnehmen.

Simulationsniveau	Chi²	Audic und Claverie	Madden
5%	gleich (0,178)	gleich (0,101)	verschieden ($p \rightarrow 0$)

Tabelle 27. Vergleich des modifizierten Z-Testes mit den anderen paarweisen Tests mit "$m \geq 5$". "Verschieden" bedeutet, daß die H_1 angenommen werden muß, "gleich", daß sie nicht angenommen werden kann. In Klammern *p*-Werte, $\alpha = 0{,}05$.

Das bedeutet, der modifizierte Z-Test sich von den anderen evaluierten Tests im Bereich (sehr) kleiner Taghäufigkeiten unterscheidet. Werden nur Tagpaare in Betracht gezogen, die eine Mindesthäufigkeit von $m \geq 5$ aufweisen, unterscheiden sich die Entscheidungen dieses Tests nur noch von dem Test nach Madden.

5.1.6 Zusammenfassung der Berechnungen zur Reliabilität von SAGE

Die globale H_0, daß die Verteilungen der beiden SAGE Profil K1 und K2 gleich ist, stellt die Grundlage für eine indirekte Überprüfung der Reliabilität von SAGE dar. Die Prüfung mittels Chi²-Test für *k* x 2- Feldertafeln (computersimulierter Verteilung der Chi²-Werte unter Gültigkeit von H_0) führte nicht zu einer Annahme der globalen H_0. Nach Bortz (1990, S. 50) lautet die Entscheidungsregel zu einer globalen H_0, daß diese bereits abgelehnt werden muß, wenn ein *einziger* Test einen statistisch bedeutsamen Unterschied aufweist. Aus diesem Grund wird in der vorliegenden Arbeit auf die Durchführung der paarweisen Tests *zum Reliabilitätsnachweis* verzichtet. Deren Ergebnis hätte keinen weiteren Einfluß auf die Gesamtaussage gehabt. Diese lautet: Die Verteilung der Taghäufigkeiten in den beiden Gruppen K1 und K2 ist statistisch also nicht gleich. Dies gilt für beide Datensätze (mit und ohne Sequenzkorrektur).

5.1.7 Sequenzfehlerkorrektur

Um den Sequenzfehler systematisch korrigieren zu können, war im Rahmen der vorliegenden Arbeit ein Verfahren entwickelt worden. Sämtliche Tests wurden auf beide Datensätze, den korrigierten und den ursprünglichen, angewandt, so daß ein Vergleich der Ergebnisse möglich ist.

Im Folgenden soll überprüft werden, ob und wenn ja, wie sich diese Korrektur auf die statistischen Ergebnisse auswirken würde. Aufgrund der Überlegung, durch die systematische Korrektur des Sequenzfehlers die Meßgenauigkeit, das heißt die Reliabilität, zu erhöhen und somit die zufällige Streuung der Werte und die Verschiedenheit der Tagpaare zu reduzieren, wurden folgende Hypothesen entwickelt:

H_0: Zwischen den beiden Datensätzen besteht bezüglich der Anzahl der Tagpaare, die auf den berechneten Niveaus einen statistisch bedeutsamen Unterschied besitzen in beiden Datensätzen, kein Unterschied.

H_1: Der Datensatz mit Sequenzfehlerkorrektur weist weniger Tagpaare auf, die auf den berechneten Niveaus einen statistisch bedeutsamen Unterschied besitzen.

Beschreibung der Vorgehensweise

Es wurde folgendes Vorgehen gewählt. Zur Überprüfung der beiden Hypothesen wurde mittels Vier-Felder-Chi²-Test die Anzahl der Tagpaare, die auf dem simulierten α-Signifikanzniveau von 5% einen statistisch bedeutsamen Unterschied aufweisen, verglichen. Die entsprechende Tafel gestaltet sich folgendermaßen:

	Daten mit Korrektur	**Daten ohne Korrektur**	**Summen**
Anzahl der Paare mit **einem statistisch signifikanten Unterschied**	a	c	N_{SIG}
Anzahl der Paare ohne **einen statistisch signifikanten Unterschied**	b	d	N_{Rest}
Summen	N_1	N_2	N_{ges}

Tabelle 28. Prototypische Vier-Felder -Tafel für den Vergleich der beiden Datensätze (mit beziehungsweise ohne Sequenzfehlerkorrektur) mittels Chi²-Test.

Der Chi²-Test überprüft, ob die Besetzung der Felder homogen ist oder statistisch signifikant verschieden. Um bei nachgewiesener Inhomogenität zu testen, welche der Felder statistisch bedeutsam über- beziehungsweise unterbesetzt sind, wurden zusätzlich die adjustierten standardisierten Residuen berechnet (analog dem Vorgehen beim Vergleich der paarweisen Tests miteinander, siehe S. 127f). Es wurde auf dem 5% Niveau getestet. Die Hypothese wurde mittels der Ergebnisse folgender Tests überprüft: Audic und Claverie (1997), Vier-Felder-Chi²-Test und Madden et al. (1997). Sämtliche Berechnungen wurden mit statistischer Software (SPSS) durchgeführt.

Aussagen zu den Hypothesen

Die H_1 kann auf dem 5% Niveau *nicht* bestätigt werden. Wenn die Anzahl der Tagpaare, die auf diesen Niveaus einen statistisch signifikanten Unterschied aufweisen, miteinander verglichen wird, ergibt sich zwischen dem Datensatz mit und demjenigen ohne Sequenzfehlerkorrektur ein statistisch bedeutsamer Unterschied (*p*-Werte siehe Tabelle 29). Dieser ist jedoch *nicht* hypothesenkonform wie den korrigierten standardisierten Residuen zu entnehmen ist. Hier ist zu sehen, daß in den Zellen des korrigierten Datensatzes, welche die Anzahl der einen statistisch signifikanten Unterschied aufweisenden Tagpaare enthalten (Zelle a in Tabelle 28), eine statistisch bedeutsame *Über*besetzung vorliegt. Die Anzahl der Tagpaare, die einen statistisch bedeutsamen Unterschied aufweisen, nimmt durch die Sequenzfehlerkorrektur also zu und nicht ab, wie in H_1 postuliert worden war.

Ergebnisse

Tabelle 29 sind die ermittelten *p*-Werte der Vier-Felder-Tafeln und Tabelle 30 die korrigierten standardisierten Residuen der Zellen, die die Anzahl der Tagpaare mit einem statistisch bedeutsamen Unterschied des Datensatzes mit Sequenzfehlerkorrektur[31] enthalten, zu entnehmen.

Audic und Claverie	4-Felder-Chi²	Madden
$3{,}7 \times 10^{-27}$	$2{,}9 \times 10^{-29}$	$8{,}2 \times 10^{-27}$

Tabelle 29. *P*-Werte des Vier-Felder-Chi²-Testes.

[31] Entspricht Zelle a der prototypischen Vier-Felder-Tafel.

Audic und Claverie	4-Felder-Chi²	Madden
10,8	11,2	10,7

Tabelle 30. Werte der korrigierten standardisierten Residuen. Es sind diejenigen Werte angegeben, die den Zellen entsprechen, die die Anzahl der Tagpaare mit einem statistisch bedeutsamen Unterschied des Datensatzes mit Sequenzfehlerkorrektur enthalten (Zelle a in Tabelle 28).

5.2 Diskussion

Die Diskussion des zweiten Teils dieser Arbeit, der sich mit statistischen Facetten von SAGE auseinandersetzt, gliedert sich folgendermaßen. Zuerst werden der statistische Entscheidungsprozeß näher erläutert (5.2.1) und Teste evaluiert, welche im Kontext von SAGE zur Anwendung kommen können (5.2.2 und 5.2.3). Anschließend werden Aspekte, die die Reliabilität betreffen, dargestellt (5.2.4).

5.2.1 Der statistische Entscheidungsprozeß

Im Folgenden sollen verschiedene Aspekte theoretischer Natur diskutiert werden, welche die Grundlagen statistischer Entscheidungen im Rahmen von SAGE bilden. Darunter fällt die Diskussion über die zu wählende Teststruktur, die Größe der Fehlerwahrscheinlichkeit, über den praktisch relevanten Unterschied sowie über den Aufbau des Entscheidungsprozesses.

5.2.1.1 Welche Teststruktur ist entscheidend?

Differenztest

Die meisten SAGE Studien werden von der Frage bestimmt, welche Unterschiede zwischen Expressionsprofilen - zum Beispiel in der Genexpression von gesundem versus pathologisch verändertem Gewebe - bestehen. Es handelt sich folglich bei der forschungsleitenden Hypothese, also bei derjenigen Hypothese, die von inhaltlichem Interesse ist, um die Alternativhypothese H_1 "Es besteht ein Unterschied.". Zur statistischen Berechnung dieser Fragestellung werden Differenztests verwandt.

Die mit dem entsprechenden Test verbundene Wahrscheinlichkeit p wird mit dem vorgegebenem Risiko α verglichen[32]. Als a priori anhand inhaltlicher Kriterien festgelegtes Signifikanzniveau gibt dieses die Wahrscheinlichkeit an, mit welcher man bereit ist, falsch

[32] α-Fehler oder Risiko 1. Art: Es wird ein Unterschied angenommen, obwohl keiner vorhanden ist.

positive Ergebnisse zu tolerieren. Liegt die ermittelte empirische Wahrscheinlichkeit p (Beobachtung|H_o)[33] unter dem Niveau dieser festgelegten α-Irrtumswahrscheinlichkeit, entscheidet man sich zugunsten der Alternativhypothese, was gleichzeitig bedeutet, daß man sich bei dieser Entscheidung zugunsten der Alternativhypothese im Rahmen der fixierten Wahrscheinlichkeit irren könnte.

Wenn die beobachtete Wahrscheinlichkeit dagegen über dem α-Niveau liegt, ist das nicht mit einer Bestätigung der H_o "Es besteht kein Unterschied." gleichzusetzen. Die Hypothese der Gleichheit kann zwar nicht abgelehnt werden, dennoch ist über das Zutreffen von H_o und H_1 bei konstanter Stichprobengröße keine Aussage möglich (Bortz 1993[4], S. 114). Wellek (1994, S.1) formuliert diesen Sachverhalt folgendermaßen: "Nichtsignifikante Unterschiedlichkeit ist nicht dasselbe wie signifikante Übereinstimmung." Dies macht deutlich, daß - soll eine H_o bestätigt werden - eine weitere Testart benötigt wird, worauf im nächsten Abschnitt eingegangen werden soll.

Äquivalenztest

Der Nachweis von Gleichwertigkeit erfordert die Anwendung eines Äquivalenztests. Analog zum Differenztest wird durch die Teststatistik die Wahrscheinlichkeit p ermittelt und mit dem vorgegebenen Risiko α verglichen. Allerdings werden die Hypothesen vorab anders formuliert. H_0 lautet " Es besteht ein Unterschied.", während H_1 - also wiederum die forschungsleitende Hypothese - "Es besteht kein bedeutsamer Unterschied." (zum Beispiel "Die Expressionsprofile sind vergleichbar.") heißt. Der weitere Entscheidungsweg entspricht grundsätzlich dem oben erläuterten. Im Rahmen von Äquivalenztests wird allerdings ein genügend enges, beidseits eingeschränktes Intervall geprüft, das die Gleichwertigkeit der zu vergleichenden Parameter charakterisiert. Es beinhaltet praktisch irrelevante, also keine bedeutsamen Abweichungen von der absoluten Gleichheit. Das bedeutet, daß im Vergleich zum Differenztest die Hypothesen nicht nur vertauscht, sondern auch modifiziert werden (Wellek 1994, S. 3). Spezifische Testverfahren finden sich bei Wellek (2003).

In der vorliegenden Arbeit wurde zum Nachweis der Reliabilität (forschungsleitende Hypothese: Gleichheit der beiden SAGE Kontrollprofile K1 und K2) kein Äquivalenztest eingesetzt, sondern ein Differenztest benutzt. Eine Aussage ist demnach nur darüber möglich,

---

[33] Lies: Die bedingte Wahrscheinlichkeit des beobachteten Ergebnisses oder extremerer unter der Annahme, daß H_o zutrifft (Bortz 1993[4], S. 110).

ob eine signifikante Verschiedenheit der beiden generierten Expressionsprofile vorliegt oder ob eine Unterschiedlichkeit der Profile, das heißt H_1 (Differenz), nicht nachzuweisen ist.

Welches Fehler*niveau* sollte gewählt werden?

Um die Frage, nach der Höhe des zu wählenden Fehlerniveaus zu beantworten, muß man sich die Gewinne und Verluste vor Augen führen, die durch die unterschiedlichen Höhen der beiden Signifikanzniveaus entstehen können. Was ist problematischer: Fehlinvestitionen (α-Fehler) oder verpaßte Gelegenheiten (ß-Fehler)?

Üblicherweise wird in biologischen Zusammenhängen, wenn es sich wie bei SAGE um eine Screeninguntersuchung handelt, für ein α-Niveau von 5% votiert. Im Gegensatz zur Wahl niedrigerer Signifikanzniveaus wird auf diese Weise der ß-Fehler auf einem relativ niedrigen Niveau gehalten. So werden weniger Unterschiede, die realiter vorhanden sind, übersehen. Die Annahme von H_1, die so gefördert wird, soll weitere Studien anregen. Gelegenheiten zu verpassen (das heißt ß-Fehler zu begehen) hieße, den Sinn der Studie zu verfehlen. Angesichts der Tatsache jedoch, daß Folgeuntersuchungen von SAGE aufwendig sind und somit eine hohe Anzahl falsch positiver Ergebnisse, also Fehlinvestitionen, die mit einem solchen hohen Signifikanzniveau verbunden sind, unerwünscht ist, scheint es angemessen, die Wahl eines niedrigeren α-Niveaus zu fordern. Diese Forderung wird von weiteren Argumenten unterstützt. Hierzu zählt, daß in den meisten Studien, auf die sich die Empfehlung eines α-Niveaus von 5% bezieht, die Stichprobengröße sehr viel geringer ist als die Anzahl der sequenzierten Tag(paare) selbst in kleinen SAGE Projekten. Dieses N liegt bei SAGE zwischen 428 Tagpaaren bei 840 Tags insgesamt im allerersten SAGE Projekt (Velculescu et al. 1995) und 49000 Tagpaaren bei 300000 sequenzierten Tags (Zhang et al. 1997) oder sogar noch höher in späteren Projekten[34]. Wie weiter unten zu sehen sein wird, ermöglicht die Erhöhung der Stichprobengröße, immer kleinere Unterschiede zwischen den Taghäufigkeiten nachzuweisen. Das heißt, daß ein "hypothesenkonformer Unterschied bei genügend großen Stichproben und einer gegebenen (endlichen) Populationsstreuung immer signifikant [wird]" (Bortz 1993[4], S. 115) und somit jede H_o bei einer genügend großen Stichprobe chancenlos ist und verworfen werden kann. Dagegen wird es bei einem festen Stichprobenumfang N schwieriger Unterschiede nachzuweisen, wenn das Signifikanzniveau immer kleiner, etwa 1% oder 0,1% gewählt wird. Wenn also bei einem kleinen Signifikanzniveau ein Unterschied als

[34] Die relevante Stichprobe ist in diesem Fall die Anzahl der ermittelten Tagpaare beziehungsweise die der sequenzierten Tags und nicht der untersuchten Individuen oder RNS Pools.

signifikant ausgewiesen werden soll, muß er hinreichend groß sein. Aufgrund der diskutierten Problematik erscheint es also empfehlenswert im Falle von SAGE ein niedriges α-Niveau zu wählen. Die Durchsicht der Literatur ergibt, daß sich die AutorInnen selten für ein α von 0,05 (beispielsweise Kal et al. 1999), sondern meist für 0,01 oder 0,001 (Madden et al. 1997, Welle et al. 2000 und andere) entscheiden.

Die folgende Möglichkeit der Datenaufbereitung findet sich in der Literatur (vergleiche zum Beispiel Audic und Claverie 1997): Anstatt das α-Signifikanzniveau *vor* der Datengewinnung aufgrund inhaltlicher Überlegungen festzulegen, kann der ermittelten *p*-Wert für jedes Tagpaar angeben werden. Im Stile eines '*Rankings*' können dann Tagpaare, deren Konstellation ein kleines *p* zur Folge hat, aufsteigend in Veröffentlichungen aufgelistet werden. Diese Vorgehensweise verwenden beispielsweise Michiels et al. (1999) und Welle et al. (2000). Audic und Claverie (1997) empfehlen es in ihrer Veröffentlichung zur Statistik von digitalen Expressionsprofilen ausdrücklich. Ein solches '*Ranking*' befreit den Untersucher oder Leser jedoch nicht von der Entscheidung, welche Transkripte weiter untersucht werden sollen, um ihre Relevanz im betrachteten Kontext nachzuweisen. Dies hat zur Folge, daß ohne die Anwendung begründeter Entscheidungskriterien eine mögliche Auswahl der interessanten Transkripte beliebig ist. Hieraus folgt, daß - sollen nachvollziehbare wissenschaftliche Entscheidungskriterien angewendet werden - die Verwendung statistischer Entscheidungskriterien letztendlich nicht vermieden werden kann. Aus diesem Grunde ist zu empfehlen, a priori ein - wie oben begründet- geringes α-Niveau zu wählen, um aus den *p*-Wert-Ranglisten mit dieser Zusatzinformation sinnvoll Gene für weitere Studien aussuchen zu können.

5.2.1.2 α-Fehler-Adjustierung

Werden auf ein und denselben Datenkörper zur Überprüfung einer globalen Hypothese Signifikanztests wiederholt angewandt, steigt die Wahrscheinlichkeit, daß mindestens einer dieser Tests fälschlicherweise signifikant wird (Bortz 1990, S. 48ff). Um das ursprünglich angestrebte α-Niveau dennoch zu halten und innerhalb des vorgegebenen Rahmens korrekt über die globale Hypothese zu entscheiden, muß eine α-Fehleradjustierung durchgeführt werden.

Im Rahmen der vorliegenden Berechnungen würde diese Problematik nur beim statistischen Vergleich der paarweisen Tests und der Evaluation der Sequenzfehlerkorrektur eine Rolle spielen. Auf eine α-Fehleradjustierung wurde verzichtet. Hierauf folgt, daß keine konfirmativen Verallgemeinerungen durchgeführt werden können. Die Datenanalyse ist

explorativ zu verstehen (mündliche Mitteilung K. Wernecke).

Im Rahmen eines üblichen SAGE Experimentes, das Unterschiede zwischen einzelnen Transkriptomen herausarbeiten möchte, ist im Gegensatz zur vorliegenden Arbeit (Aspekt Reliabilitätsprüfung) die globale Hypothese, also der Vergleich der Gesamtverteilungen sekundär[35]. Es interessiert inhaltlich nicht, ob zwischen den beiden zu vergleichenden Expressionsprofilen *insgesamt* ein Unterschied besteht, sondern nur, welche einzelnen Transkripte reguliert erscheinen. Dennoch sind die verschiedenen Tagpaare über die Gesamttagzahl miteinander verbunden, was für die Durchführung einer Adjustierung sprechen würde. Außerdem können manche Transkripte einer funktionellen Einheit angehören, wodurch ihre Regulation interdependent ist.

Im Folgenden soll dargestellt werden, welche Möglichkeiten zur α-Fehlerkorrektur im Rahmen von üblichen SAGE Experimenten sinnvoll anzuwenden wären.

Eine traditionelle Form der Adjustierung ist die sogenannte Bonferroni-Korrektur. Hierbei errechnet sich α_{KORR}, das α-Niveau, das für den einzelnen Test entscheidend ist, nach: $\alpha_{KORR} = \alpha / k$, wobei k die Anzahl der durchgeführten Tests ist (Bortz 1990, S.51). α_{KORR} wäre zum Beispiel auf die vorliegenden Daten angewandt angesichts der 14159 Paare (Datensatz ohne Sequenzfehlerkorrektur) sehr klein (beispielsweise für ein globales α von 5%: 0,00035%). Das heißt, das diese Art der Korrektur äußerst konservativ ist und damit einem Übersichtsverfahren wie SAGE nicht angemessen.[36] Etwas progressiver wäre eine sequentielle Weiterentwicklung der Bonferroni-Korrektur (Holm 1979). Hierzu werden die Tagpaare nach Bestimmung ihres jeweiligen empirischen P-Wertes aufsteigend angeordnet; das heißt, beginnend mit dem Paar, das den kleinsten P-Wert aufweist. Über die Signifikanz des ersten, das heißt des kleinsten Wertes wird anhand eines α_{KORR} entschieden, das entsprechend der Gleichung für die Bonferroni-Korrektur bestimmt wurde. Das nächst größere P wird mit einem zweiten α_{KORR} verglichen, das nach $\alpha_{KORR} = \alpha / (k\text{-}1)$ berechnet wurde. Dies wird solange wiederholt bis $p_n > \alpha / (k\text{-}n)$ ist, so daß dieser und alle folgenden Vergleiche keine signifikanten Unterschiede auf dem gewählten Niveau mehr aufweisen. Doch auch diese sequentielle Vorgehensweise führt dazu, daß nur sehr wenige Gene als reguliert erkannt werden (Yang 2003, S. 60f). Beide Verfahren kontrollieren die FWER

---

[35] Auf die exakte Vorgehensweise statistischer Beweisführung im Rahmen eines üblichen SAGE Experimentes (1. Überprüfung der globalen Hypothesen, 2. paarweise Signifikanzberechnung) wird weiter unten (siehe S. 134) eingegangen.

[36] Im Falle des Test von Audic und Claverie beispielsweise wäre erst ab 497 Tagpaaren (damit α_{KORR} = 0,00050302) oder weniger anstelle von 14159 Tagpaaren (Daten ohne Sequenzfehlerkorrektur) H_1 anzunehmen.

('family-wise error rate'), welche die Wahrscheinlichkeit angibt, mindestens einen Typ I-Fehler zu begehen.

In den letzten 10 Jahren wurden andere Ansätze zur Fehlerkontrolle entwickelt, die insbesondere im Rahmen der Mikroarray- und Chiptechnologien zur Anwendung kommen. Hier ist die 'false discovery rate' (FDR) hervorzuheben (Benjamini 1995, Weiterentwicklung beispielsweise Storey 2002: positive FDR). Diese ist der erwartete Anteil an Typ I- Fehlern unter den abgelehnten Hypothesen, also den als reguliert deklarierten Genen. Der Vorteil der darauf aufbauenden Verfahren liegt darin begründet, daß der α-Fehler streng versuchsbezogen kontrolliert wird und damit nicht ins Unermeßliche steigen kann, und auf der anderen Seite genügend Gene als reguliert identifiziert werden, die für weitere Untersuchungen relevant sein können.

Ein wichtiger Gesichtspunkt bei der Entwicklung dieser Verfahren ist die Abhängigkeit der Einzeltests untereinander. Die Korrekturform nach Bonferroni beispielsweise trägt diesem nicht Rechnung (Dudoit 2003). Dudoit (2003) nennt einige statistische Verfahren, welche auf diese Problematik Bezug nehmen (unter anderem FDR-Verfahren, siehe hierzu auch Storey 2002).

Ein weiterer Aspekt ist die sogenannte "Stärke der Fehlerkontrolle": Eine starke Prüfung des α-Fehlers kontrolliert diesen unabhängig von der Kombination von regulierten und konstant expremierten Genen, während eine schwache Kontrolle davon ausgeht, daß kein Gen reguliert ist, was im Rahmen von SAGE unwahrscheinlich ist, so daß Verfahren mit starker Kontrolle zu bevorzugen sind. Hierzu zählen unter anderem die Bonferroni-Methode, die FDR nach Benjamini 1995 und deren Weiterentwicklungen (Dudoit 2003).

Auch bezüglich der FWER gibt es Weiterentwicklungen, die den Erfordernissen von molekularbiologischen Verfahren mit tausenden von zu testenden Einzelhypothesen Rechnung tragen (Westfall 1993). Hierunter scheint jedoch nur das maxT Verfahren sinnvoll anwendbar zu sein (Yang 2003).

Im Rahmen von üblichen SAGE Experimenten werden moderne Verfahren bisher selten abgewandt. Beispiele für eine α-Fehler-Kontrolle mittels FDR wären Hosack 2003, Hauser 2003 und Divina 2004, wobei hier nur die ursprüngliche Form der FDR (Benjamini 1995) verwandt wird (Ausnahme Divina 2004).

Die statistischen Resultate werden zusätzlich zum α-Niveau durch drei weitere Kriterien bei der Entscheidung, welche Transkripte weiter untersucht werden sollen, unterstützt: Einmal spielt es eine Rolle, welcher funktionellen Gruppe die fraglichen Transkripte angehören

(Carulli et al. 1998). Bestimmte Bereiche mögen aufgrund der bereits vorhandenen Literatur interessanter erscheinen als andere. Dieses Kriterium ist also inhaltlicher Natur und muß im Kontext der jeweiligen Untersuchung diskutiert werden. Die anderen beiden Kriterien beziehen sich dagegen auf die Struktur der Daten. Es handelt sich dabei um die Häufigkeiten der Tags, und um die Höhe ihres Unterschieds (so er auf dem gewählten Niveau signifikant ist). Im folgenden Abschnitt sollen diese beiden Kriterien diskutiert werden.

5.2.1.3 Praktische Bedeutsamkeit

Regulation

Mit Erhöhung der Stichprobengröße ins Unendliche kann theoretisch jeder noch so kleine Unterschied statistisch bedeutsam, das heißt signifikant sein. Das bedeutet, daß für die Entscheidung, welche Transkripte in weiteren Studien betrachtet werden sollen, nicht nur erheblich ist, ob sie in statistischen Tests einen signifikanten Unterschied aufweisen, sondern auch wie groß dieser ist[37]. Transkripte, welche beispielsweise lediglich um ein Drittel reguliert[38] erscheinen, sind es unter Umständen nicht wert weiteruntersucht zu werden, auch wenn kleine Veränderungen der Genexpression in biologischen Systemen eine große Wirkung haben können. Für die Praxis von SAGE und den nachfolgenden (Funktions-)Untersuchungen entscheiden sich die meisten Autoren für eine mindestens zweifache Regulation (zum Beispiel Kal et al. 1999, Angelastro 2000a). Manche Publikationen konzentrieren sich jedoch sogar nur auf Transkripte, welche mindestens fünffach (Lal et al. 1999), zehnfach (Yu et al. 1999) oder sogar zwölffach (Hashimoto et al. 1999) reguliert erscheinen. Diesem Konzept entspricht in der Statistik die Effektgröße ε, die aufgrund inhaltlicher Erwägungen festgelegt wird, um einen praktisch bedeutsamen Unterschied zwischen Parametern von H_o und H_1 zu definieren (Bortz 1993[4]). Gerade, wenn zukünftig Sequenzierverfahren noch effizienter werden und die Produktion einer sehr großen Menge Tags auch in Standardlabors möglich ist, ist es wichtig, einen solchen praktisch relevanten Unterschied festzulegen, da sonst undeutlich wird, welche der statistisch signifikanten Unterschiede der Genexpression es sich lohnt, weiter zu verfolgen.

Darüber hinaus werden Stichprobenumfänge dann als optimal bezeichnet, wenn bei

---

[37] Die "statistische Signifikanz [ist] eine notwendige, aber keine hinreichende Bedingung für praktische Bedeutsamkeit." (Bortz 1990, S. 42).

[38] Lal et al. (1999) weisen darauf hin, daß die beiden zu vergleichenden Profile vorher normalisiert werden müssen. Zur Berechnung der Größe des Unterschieds zwischen den beiden Häufigkeiten eines Tagpaars siehe S. 49.

festgelegtem α und ß ein bestimmter erwarteter Unterschied nachgewiesen werden kann. Der mit der Sequenzierung einer großen Menge Tags verbundene Aufwand lohnt sich dann nicht, wenn eine unter praktischen Gesichtpunkten für bedeutsam erachtete Effektgröße auch mit einem kleineren, dem 'optimalen' Stichprobenumfang, abgesichert werden könnte (Bortz 1993[4], S. 120). Die Software SAGEstat bietet die Möglichkeit zu derartigen Kalkulationen.

Minimale Taghäufigkeiten

Einer anderer Aspekt der praktischen Bedeutsamkeit sind die Häufigkeiten, die Tags mindestens aufweisen müssen, um als relevant erachtet zu werden.
Man et al. (2000) empfehlen nur Tags näher zu betrachten, die öfter als 10 mal auftreten. Welle et al. (2000) konzentrieren ihre Auswertung ebenfalls auf Tags, die mindestens 10 mal in einem der beiden Expressionsprofile präsent sind. Während beispielsweise Michiels et al. (1999) oder Larson et al. (2000) schon Tags in ihre Analyse einbeziehen, die 5 mal und Angelastro et al. (2000a) sogar solche, die nur 2 mal vorhanden sind. Dabei ist zu beachten: Eine bestimmte Tagmenge bedeutet im Kontext einer kleinen Gesamtmenge etwas anderes als bei einer sehr großen Menge an sequenzierten Tags, so daß das Kriterium "minimale Taghäufigkeiten" im Rahmen des jeweiligen Projektes festgelegt werden sollte.
Ein wichtiger Gesichtspunkt ist zudem der Kontext des jeweiligen Projektes. Während es bei zerebralen Expressionsprofilen aufgrund der Komplexität des Expressionsmusters sinnvoll sein kann, auch Transkripte zu betrachten, welche sehr gering exprimiert erscheinen, ist bei Geweben geringer Komplexität das Gegenteil der Fall. Sinnvoll wäre es, nicht einzelne Taghäufigkeiten als Grenzwert zu bestimmen (wie Man et al. 2000 es zum Beispiel machen), sondern das gemeinsame arithmetische Mittel oder die Summe zweier zu vergleichender Transkripte als Richtlinie zu nehmen wie es in der vorliegenden Arbeit gehandhabt wurde.
Auf das Thema der geringen Häufigkeiten wird nochmals unter dem Gesichtpunkt der Reliabilität eingegangen werden (siehe S. 173).

5.2.1.4 Struktur des statistischen Entscheidungsprozesses

Wie bereits deutlich geworden ist, können die Daten, die SAGE produziert, unter zwei verschiedenen Blickwinkeln betrachtet werden. Einmal kann die Frage nach dem Zusammenhang zweier SAGE Profile beziehungsweise nach deren Gleichheit oder Unterschiedlichkeit *global* für die Gesamtverteilung der Profile gestellt werden, und einmal *individuell* für jedes einzelnen Tagpaar.
Bei der quantitativen Auswertung und dem Vergleich zweier SAGE Profile stellt sich als

erstes die Frage, ob die beiden Datenreihen insgesamt als homogen anzusehen sind oder nicht. Erst wenn hier ein geeigneter Test (zum Beispiel der in der vorliegenden Arbeit verwendete Chi²-Test mit Iterationen) diese globalen Hypothesen evaluiert hat und die Verteilungen als inhomogen angenommen werden können, kann in Einzelvergleichen untersucht werden, welche Tagpaare diese Inhomogenität verursachen. Bezüglich ihrer inhaltlichen Relevanz unterscheiden sich diese beiden Blickwinkel, je nachdem, ob es sich um eine Reliabilitätsstudie wie die der vorliegenden Arbeit oder aber um ein übliches SAGE Experiment handelt. In der vorliegenden Arbeit ist die Prüfung der globalen H_0 inhaltlich bedeutsam, während üblicherweise nur der direkte Vergleich der einzelnen Transkripte und ihres Expressionsniveaus von Interesse ist. Dennoch sollte diese inhaltliche Präferenz das statistische Vorgehen bei üblichen Experimenten nicht dominieren. Es sind in der Literatur jedoch nur zwei Veröffentlichungen bekannt, die dieser korrekten Struktur des statistischen Entscheidungsprozeß folgen (Michiels et al 1999 und Margulies et al. 2001).

Nachdem in den oben stehenden Abschnitten die wesentlichen Aspekte des statistischen Entscheidungsprozesses eines SAGE Projektes deutlich geworden sind, stellt sich die Frage, welche der paarweisen Test unter welchen Bedingungen zur Anwendung kommen sollten. Dies soll in den nun folgenden beiden Kapiteln 5.2.2 und 5.2.3 diskutiert werden.

5.2.2 Evaluation nicht angewandter Tests

5.2.2.1 *Bayes-Test nach Chen et al. (1998)*

Beschreibung

Diese Methode wurde von Chen et al. (1998) entwickelt und von Lal et al. (1999) und Lash et al. (2000) für den Gebrauch einer interaktiven Webseite ("SAGEmap") modifiziert. Der Ansatz berechnet die Posterior-Wahrscheinlichkeit, daß das Expressionsniveau eines Transkriptes mindestens um einen bestimmten (vorher festzulegenden) Faktor angestiegen ist. Dazu muß die Verteilung der Prior-Wahrscheinlichkeit (vor der Beobachtung der Taghäufigkeiten) abgeschätzt werden. Auf eine Darstellung der mathematischen Herleitung soll an dieser Stelle verzichtet werden. Mittels der Internetseite SAGEmap (www.sagenet.org) kann für dort veröffentlichte oder hochgeladene eigene SAGE Profile diese Posterior-Wahrscheinlichkeit nach Wahl des kleinsten Regulationsniveaus, das als relevant erachtet wird, berechnet werden.

Diskussion des statistischen Ansatzes

Der Ansatz nach Chen et al. (1998) weist einige Einschränkungen auf.
Chens statistisches Verfahren setzt annähernd gleiche Gesamttagmengen voraus, was seine Anwendungsmöglichkeit erheblich reduziert, wenn nicht sogar unmöglich macht (siehe S. 146). Die Modifikation von Lal et al. (1999) läßt jedoch auch unterschiedlich Gesamtmengen zum Vergleich zu. Für die Wahl der Größe der Parameter der Prior-Wahrscheinlichkeit gibt es keinen verbindlichen Wert, sie unterscheiden sich je nach Gewebe, Krankheit etc. Dadurch können die Ergebnisse des statistischen Entscheidungsprozesses inkonsistent werden (Man et al. 2000). Unklar ist, ob die fehlende Anpassung der Parameter (wie es auf der Webseite der Fall ist) nicht sogar zu falschen Ergebnissen führen kann. Chen et al. (1998) berufen sich bei der Abschätzung dieser Parameter nur auf die Verteilung ihrer eigenen Daten, Lal et al. (1999) beziehen zwei weitere SAGE Projekte (Zhang et al. 1997 und Polyak et al. 1997) mit ein. Der Ansatz nach Chen et al. (1998) berechnet die Wahrscheinlichkeit für die Gültigkeit folgender Hypothese: Das Expressionsniveau eine bestimmten Transkriptes ist um einen Faktor k (im Voraus festgelegt) angestiegen. Die Hypothese, die diesem statistischen Test zugrunde liegt, unterscheidet sich damit von den in dieser Arbeit überprüften Hypothesen. Sie ist spezifisch und gerichtet, was einem einseitigen Test entspricht, während die in dieser Arbeit vorgestellten Hypothesen unspezifisch und ungerichtet sind. Zu beachten ist hierbei, daß eine gerichtete und spezifische Hypothese bereits durch geringere Differenzen bestätigt wird als eine ungerichtete und unspezifische (Bortz 1993[4], S. 114). Der hier vorgestellte Test weist somit geringere Häufigkeitsunterschiede zwischen Tagpaaren als statistisch bedeutsam aus als Tests, die unspezifische und ungerichtete Hypothesen prüfen. Es werden damit mehr Tagpaare statistisch signifikant. Zu beachten ist außerdem, daß der Test zweimal durchgeführt werden muß, um auch Transkripte zu erfassen, die herunterreguliert sind. Daraus folgt, daß eine α-Fehleradjustierung vorgenommen werden sollte. Aufgrund der genannten Einschränkungen des Tests wurde in der vorliegenden Arbeit auf dessen Durchführung verzichtet.

5.2.2.2 Fishers Exakt Test

Beschreibung

Dieser Test wird üblicherweise als Alternative zum Vier-Felder-Chi²-Test eingesetzt, wenn es sich um Tafeln mit sehr kleinen Besetzungszahlen handelt (Sachs 1999[9], S. 477). Der Test basiert auf der hypergeometrischen Verteilung und fragt bei fixierten Randsummen, "nach der

Wahrscheinlichkeit dafür, daß die beobachtete Besetzung der Tafel oder eine noch weniger wahrscheinliche rein zufällig zustanden kommt" (Sachs 1999[9], S. 477).

Diskussion des statistischen Ansatzes

Fishers Exakt Test ist als konservativ entscheidend bekannt (Sachs 1999[9], S.477). Das heißt, daß weniger Werte signifikant werden, als nach dem nominellen α-Niveau zu erwarten wäre. So berichten auch Man et al. (2000) in ihrer bereits erwähnten Monte-Carlo-Studie zur Teststärke und Robustheit von Tests, die im Kontext von SAGE zur Anwendung kommen können, daß Fishers Exakt Test in Vergleich zum Vier-Felder-Chi²-Test eine geringere Teststärke[39] und Robustheit[40] besitzt. Dies hat zur Folge, daß das Risiko einen ß-Fehler[41] zu begehen höher ist. Der Test ist zudem für die großen Zahlen, die ein SAGE Projekt wie das vorliegende beinhalten, extrem aufwendig zu berechnen.
Aus diesen Gründen wurde in der vorliegenden Arbeit auf die Anwendung dieses Tests verzichtet.

Verwendung in der Literatur

Obwohl der Test als der vielleicht am meisten akzeptierte für Vier-Felder-Tafel bezeichnet wird und im Rahmen von übergreifenden EST Projekten zur Genexpressionsanalyse als Standard vorgeschlagen wird (www.ncbi.nlm.nih.gov/UniGene/fisher.shtml, 1.10.2001), scheint er sich im Kontext von SAGE nicht durchzusetzen. So ergibt die Literaturrecherche (Stand 2002) kaum Publikation, die Fishers Exakt Test zur statistischen Analyse von SAGE Daten verwendet. Ein Beispiel wäre Trendelenburg et al. 2002.

5.2.2.3 SAGE 300

Beschreibung

Die von Zhang et al. (1997) vorgestellte Software zur Auswertung von SAGE-Daten enthält unter anderem Möglichkeiten, diese statistisch zu analysieren. Der verwendete Ansatz beruht auf einer Monte-Carlo-Studie ohne auf spezifische statistische Testverfahren zurückzugreifen. Es wird dabei die relative Wahrscheinlichkeit ermittelt, daß - bei Gültigkeit der H_0[42] - die

---

[39] Die Teststärke gibt an, mit welcher Wahrscheinlichkeit ein Test zugunsten einer Alternativhypothese entscheidet (Bortz 1993[4], S. 118): Teststärke = 1 - ß

[40] Nach Bortz (1990, S. 83) bezeichnet die Robustheit eines statistischen Testes seine "Unempfindlichkeit [...] gegenüber Voraussetzungsverletzungen und gegenüber ungewöhnlichen Stichprobencharakteristika.".

[41] ß-Fehler: Die H_0 wird angenommen, obwohl die H_1 gilt.

[42] Als H_0 gilt, daß das Niveau, die Art und die Verteilung der Transkripte in den beiden zu vergleichenden

beobachtete Differenz (oder eine größere) zwischen den beiden Häufigkeiten eines Tagpaares aufgrund zufälliger Schwankungen zustande gekommen ist. Dazu werden 100000 Simulationszyklen durchgeführt[43]. Der Wert der resultierenden Wahrscheinlichkeit repräsentiert denjenigen Anteil der Simulationsergebnisse, der die beobachtete Differenz oder eine größere aufweist. Um diese relative Wahrscheinlichkeit in eine absolute zu konvertieren, werden 40 Experimente simuliert, in welchen eine repräsentative Anzahl von Transkripten identifiziert und verglichen wird. Die Verteilung der dazu verwendeten Transkripte wird vom durchschnittlichen Expressionsniveau, wie es den experimentell beobachteten SAGE Profilen zu entnehmen ist, abgeleitet. Die relativen Wahrscheinlichkeiten, die in diesen 40 simulierten Experimenten ermittelt werden, entsprechen falsch positiven Ergebnissen (α-Fehler). Die Verteilung dieser *p*-Werte wird mit der Verteilung verglichen, welche die anhand der experimentellen Daten simulierten relativen Wahrscheinlichkeiten wiedergibt. Anhand dieses Vergleichs kann - im Sinne einer α-Fehlerkorrektur - der maximale *p*-Wert festgelegt werden, der (entsprechend der Hypothesen) einen möglichen Nachweis von Unterschieden in der Genexpression mit der zuvor gewählten Irrtumswahrscheinlichkeit sichert[44].

Diskussion des statistischen Ansatzes

Die statistische Entscheidungsfindung der SAGE 300 Software orientiert sich an den beobachteten Daten beziehungsweise den darauf aufbauenden Computersimulationen. Es wird keine Teststatistik verwendet, die auf einem bestimmten mathematischen Modell basiert, wodurch die Anzahl der zugrundeliegenden Annahmen minimiert wird und keinerlei Einschränkungen bezüglich der Taghäufigkeiten bestehen. Dieser Aufbau des Tests hat allerdings zur Folge, daß sich der simulierte *p* - Wert, den die Software ermittelt, bei jedem Testdurchlauf ändert (auch wenn die Eingabe identisch ist). Ein exakter Vergleich dieses Tests mit denjenigen, die den *p* - Wert exakt errechnen, gestaltet sich deswegen schwierig, so daß auf die Durchführung dieses Tests in der vorliegenden Arbeit verzichtet wurde. Ruijter et al. (2002) vergleichen SAGE300 dennoch mit anderen paarweise prüfenden Tests (SAGEstat, Fishers Exakt Test, Test nach Madden und Test nach Audic und Claverie) und kommen zu dem Schluß, daß die Ergebnisse der Simulationen von SAGE300 mit von denjenigen von SAGEstat, Fishers Exakt Test und dem Test nach Audic und Claverie (1997) übereinstimmen.

Populationen gleich ist.

[43] Laut Ruijter (1999) werden die Simulationen beendet, wenn 100 Zyklen zu einer Differenz geführt haben, die genauso groß (oder größer) wie die beobachtete ist.

[44] Das Testresultat ist einseitig (Ruijter 1999).

5.2.3 Evaluation der angewandten Tests

Im Folgenden sollen Tests diskutiert werden, die im Rahmen der vorliegenden Arbeit zur Anwendung kamen beziehungsweise geprüft wurden. Es wird dabei auf Besonderheiten der statistischen Ansätze, gegebenenfalls auf die Voraussetzungen zur gültigen Anwendung, die Praktikabilität und die Verwendung in der Literatur eingegangen werden.

5.2.3.1 Tests zum Vergleich der Gesamtverteilungen

5.2.3.1.1 Chi²-Test für *k* x 2- Felder Tafel (Simulationen)

Diskussion des statistischen Ansatzes

Anhand von Zufallszahlen und Iterationen eine Verteilung der Chi²-Werte bei Gültigkeit von H_0 zu erstellen, ist eine gute Möglichkeit, um angesichts der Tatsache, daß die Voraussetzungen zur Anwendung des *k* x 2- Felder-Chi²-Testes nur eingeschränkt erfüllt werden können, eine Kennwertverteilung für H_0 zu erzeugen, die sich an den jeweils konkret vorliegenden Daten orientiert. Der *p*-Wert der beobachteten Verteilung kann entsprechende der Formel für den *k* x 2- Felder-Chi²-Test exakt berechnet werden.

Verwendung in der Literatur

Im Rahmen üblicher SAGE Experimente ist es sinnvoll wie oben dargestellt, vor der paarweisen Überprüfung der Profile zu untersuchen, ob insgesamt betrachtet ein Unterschied zwischen den Verteilungen vorliegt. Dieser Ansatz wurde von Ruijter (1999) im Rahmen des Vortrags "SAGE and Statistics" (SAGE Workshop, Hilversum) vorgestellt. Michiels et al. (1999) und Margulies et al. (2001) folgen der Vorgehensweise.

5.2.3.1.2 Kontingenzkoeffizient

Diskussion des statistischen Ansatzes

Um das Assoziationsmaß von Kontingenztafeln auszudrücken, gibt es diverse Koeffizienten: Phi, Pearsons C, Cramérs V und andere. Cramérs V wird im allgemeinen bevorzugt (Rasch 1996, S. 617), da dieser Koeffizient im Gegensatz zu anderen unabhängig von der Größe der Kontingenztafel zwischen 0 (totale Unabhängigkeit) und 1 (totale Abhängigkeit) liegt. Aus diesem Grund wurde er auch in der vorliegenden Arbeit zur Berechnung des Assoziationsmaßes gewählt.

Folgende Einschränkungen sind bei der Interpretation der vorliegenden Ergebnisse zu berücksichtigen. Erstens geht die Tatsache, daß N_1 ungleich N_2 ist, insofern nicht in die Berechnungen ein, als durch den Aufbau der Kontingenztafel auch eine Taghäufigkeit von

Null als Beobachtung gilt, wenn das Tag in der anderen Gruppe gefunden wurde.

Ein zweiter Aspekt ist die Frage nach der Erfüllung der Voraussetzungen des χ^2-Testes (siehe auch S. 113), insbesondere der Größe der erwarteten Häufigkeiten. In der Regel wird davon ausgegangen, daß sämtliche Erwartungshäufigkeiten mindestens den Wert fünf annehmen sollten. Aufgrund der heterogenen Verteilung der Daten ist dies bei den betrachteten Kontingenztafeln nicht immer gegeben. Zur Verwendung des Testes im Zusammenhang mit der Reliabilitätsprüfung siehe S. 163).

5.2.3.2 Tests zum paarweisen Vergleich

Um neben den theoretischen Überlegungen und der Diskussion der vorhandenen Literatur die Tests konkret miteinander vergleichen zu können, wurde die statistische Analyse „normaler" SAGE Experimente beispielhaft nachgeahmt ($\alpha = 5\%$). Die Ergebnisse dieser Berechnungen wurden statistisch evaluiert. Aufgrund der Komplexität der Testsituation sind diese Ergebnisse bei fehlender α-Fehleradjustierung explorativ zu verstehen.

5.2.3.2.1 Test nach Madden et al. (1997)

Statistischer Ansatz

Ein Vorteil des Tests nach Madden et al. (1997) ist, daß er sehr einfach mit einem Tabellenkalkulationsprogramm zu berechnen ist. Der Hauptkritikpunkt an diesem Ansatz liegt darin begründet, daß die Gesamthäufigkeiten der zu vergleichenden Expressionsprofile nicht berücksichtigt werden. Dies hat zur Folge, daß dieser Test genaugenommen lediglich zum Vergleich zweier SAGE Bibliotheken gleicher Größe angewendet werden kann (Kal et al. 1999 und Ruijter 1999)[45]. Dies ist bei SAGE jedoch selten der Fall. Selbst wenn die beiden zu vergleichenden Profile im Rahmen eines einzigen Projektes erstellt werden und angestrebt wird, in beiden Gruppen die gleiche Menge Tags zu sequenzieren, können sich diese Zahlen aufgrund der Auswertung (Elimination von Linkerartefakten, redundanten Dimeren und Korrektur des Sequenzfehlers) anschließend wesentlich verändern. Es ist folglich realistischer Weise nicht möglich, davon auszugehen, daß die für diesen Test erforderliche Bedingung im Normalfall erfüllt werden kann.

---

[45] Die in dieser Arbeit verglichenen Gruppen unterscheiden sich nur um 2,38 % in ihrer Gesamtzahl voneinander. Der Test nach Madden et al. (1997) wurde unter der Annahme, daß diese Differenz zu vernachlässigen sei, durchgeführt.

Kal et al. (1999) weisen darauf hin, daß der Test nach Madden et al. (1999) konservativ entscheidet. Das heißt, daß die faktische α-Fehlerwahrscheinlichkeit[46] unter dem festgelegten Signifikanzniveau liegt. Es werden dadurch weniger Tagpaare als statistisch signifikant verschieden ermittelt, als das mit einem weniger konservativ entscheidenden Test der Fall wäre (siehe auch S. 154). Ruijter (2002) stellt einen Vergleich verschiedener Test (SAGE 300 Software, Vier-Felder-Chi²-Test, nicht modifizierter Z-Test (SAGEstat), Audics und Claveries Test, Maddens Test) an. Analog zum Vorgehen der vorliegenden Arbeit werden die Häufigkeitswerte verglichen, bei welchen die verschiedenen Tests aus dem 5% Niveau einen statistisch bedeutsamen Unterschied angeben. Der Test nach Madden et al. (1997) unterscheidet sich dem Augenschein nach deutlich von den anderen, die einheitlich zu entscheiden scheinen. Im Gegensatz zu den hier vorliegenden Ergebnisse werden die Befunde jedoch nicht statistisch abgesichert. Die Untersuchung der vorliegenden Arbeit kann diese Unterschiedlichkeit des Tests von Madden et al. (1997) nicht bestätigen.

Verwendung in der Literatur

Der Test nach Madden et al. (1997) wird selten verwendet. Ein Beispiel wäre Hashimoto et al. (1999).

5.2.3.2.2 Test nach Audic und Claverie (1997)

Voraussetzung: Poissonverteilung

Die mathematische Grundlage der von Audic und Claverie (1997) entwickelten Statistik ist die Annahme, daß Transkripte poissonverteilt sind. Diese Annahme wurde an den vorliegenden Daten mittels Kolmogorov-Smirnov-Anpassungstest (SPSS Version 10.0) auf dem 1% Niveau überprüft. Dieser Test vergleicht die beobachtete kumulierte mit der theoretischen Verteilung. Bei *z*-Werten von 15,1 (K1 ohne Korrektur), 15,7 (K2 ohne Korrektur), 18,1 (K1 mit Korrektur) und 18,8 (K2 mit Korrektur) wurde in allen vier Fällen *p*-Werte $\rightarrow 0$ erreicht. Das heißt, daß die H_0 (Die beobachtete Verteilung ist poissonverteilt) auf dem 1% Niveau abgelehnt werden muß. Die Daten sind also nicht poissonverteilt. Dies hat zur Folge, daß die Voraussetzung des Testes wahrscheinlich nicht erfüllt sind. Allerdings muß das nicht heißen, daß der Test bei den vorliegenden Daten nicht angewandt werden darf. Möglich wäre die Beantwortung der Frage, wie der Test auf diese Verletzung seiner Vorraussetzung reagiert, per Monte Carlo Studie (Bortz 1993[4], S.125). Alternativ könnte ein

[46] α-Fehler: Die H_1 wird angenommen, obwohl die H_0 gilt.

verteilungsfreier Test durchgeführt werden, der an weniger Voraussetzungen geknüpft ist, jedoch eine geringere Teststärke (T = 1 - ß) besitzt (Bortz 1993[4], S.125). Mit dem Vier-Felder-Chi²-Test wird ein derartiger Test vorgestellt.

Statistischer Vergleich mit anderen Tests

Die Ergebnisse der Simulationen üblicher SAGE Experimente dienten dem statistischen Vergleich der Tests. Die Anzahl der als signifikant verschieden ermittelten Tagpaare unterscheiden sich statistisch nicht von derjenigen des Test nach Madden et al. (1997). Ein Unterschied besteht beim Vergleich mit dem Chi²- Test sowie dem modifizierten Z-Test (siehe S. 129f, S. 151 und S. 154).

Testrobustheit

Nach Bortz (1990, S. 83) bezeichnet die Robustheit eines statistischen Testes seine "Unempfindlichkeit [...] gegenüber Voraussetzungsverletzungen und gegenüber ungewöhnlichen Stichprobencharakteristika.". Analysen dieser Eigenschaft werden meist im Rahmen von Monte-Carlo-Studien durchgeführt. Mittels beispielsweise tausendfachem Ziehen von Zufallsstichproben aus einer Population von Zufallszahlen, für welche H_o gilt, wird hierbei die Rate an richtigen und falschen Entscheidungen des Testes berechnet.

Audic und Claverie berichten in der Veröffentlichung von 1997 zu dem von ihnen entwickelten Signifikanztest von einer Monte-Carlo-Studie zur Überprüfung der Robustheit ihres Test in Abhängigkeit von verschiedenen Expressionsniveaus. Der Test zeige bei einer Taghäufigkeit von $n < 5$ ein leicht konservatives Verhalten. Das heißt, daß die Rate an falsch positiven Resultaten unter dem gewählten α-Niveau liegt. Bortz (1993[4], S. 125) bewertet eine solche Reaktion als akzeptabel, wenn man mit einer reduzierten Rate an signifikanten Ergebnissen einverstanden ist. Bei Taghäufigkeiten von $n \geq 5$ erweist sich der Test von Audic und Claverie als robust. Das Niveau von α wird knapp erreicht, jedoch nicht überschritten.

Es liegen Ergebnisse von Man et al. (2000) vor, die in diesem Kontext von besonderem Interessen sind. Die Autoren führten per Monte-Carlo-Studie einen Vergleich der Robustheit dreier statistischer SAGE Test bezüglich des α- und ß- Fehlerniveaus durch (paarweiser Chi²-Test, paarweiser Fishers Exakt Test und Audics und Claveries Ansatz). An dieser Stelle soll nur von den Ergebnissen zu Audics und Claveries Test berichtet werden. Für einheitliche Gesamttagzahlen (N=50000 in beiden Stichproben) ergeben sich im Falle von Taghäufigkeiten $n \geq 20$ ein durchgehend kleiner ß-Fehler (circa 1%) sowie ein konstantes α (nahe 5%). Sobald die Taghäufigkeiten n unter 15 sinken, steigt der ß-Fehler rapide an (bis

auf fast 100%), während α gleichzeitig abnimmt. Wenn unterschiedliche Taggesamtzahlen eingesetzt werden (50000 versus 250000 und vice versa), scheint der Test besonders wenig robust zu sein. Man et al. (2000) spezifizieren die entsprechenden Resultate nur für den Bereich kleiner Häufigkeiten: Bei Häufigkeiten von $n = 15$ ergibt sich ein ß-Fehler von circa 45%, der kontinuierlich ansteigt, wenn die Taghäufigkeiten Richtung $n = 1$ gehen. Der α-Fehler bleibt dagegen im Bereich von $n = [15; 5]$ Tags konstant, um erst darunter abzunehmen. Die Ergebnisse zum α-Fehler stimmen mit der Beobachtung von Audic und Claverie (1997) überein. Auch bei den Ergebnissen der vorliegenden Arbeit läßt sich beobachten, daß die Rate der Tagpaare, die einen statistisch bedeutsamen Unterschied aufweisen, höher ist, wenn nur Tagpaare mit einem Mittelwert von $m \geq 5$ untersucht werden (zum Beispiel 13,6% aller Paare), als die Rate, die Tags aufweisen, die einen Mittelwert von $m < 5$ haben (gleicher Fall 1,6%).

Welche Auswirkungen haben die Ergebnisse der Monte-Carlo-Studien für den experimentellen Vergleich zweier SAGE Profile?

Um diese Frage zu beantworten, muß der α-Fehler betrachtet werden. Dieser weicht erst ab einer Taghäufigkeit von $n < 5$ vom vorgegebenen Niveau nach unten ab. Im Bereich derartig kleiner Taghäufigkeiten liegt die de facto Irrtumswahrscheinlichkeit somit unter der gewählten. Diese Eigenschaft macht den Test von Audic und Claverie (1997) zu einem geeigneten Test, wenn keine minimale Taghäufigkeit festlegt werden soll (wie es zum Beispiel in SAGE Projekten mit einer vergleichsweise geringen Anzahl an sequenzierten Tags der Fall sei könnte), aber in diesem unteren Bereich, die Daten unter strengen Kriterien betrachtet und wenig falsch positive Transkripte riskiert werden sollen.

Anwendung

Ohne eine Automatisierung der Berechnungen sind diese mit einem Tabellen-kalkulationsprogramm wie zum Beispiel Excel kaum praktikabel. Auf einer Internetseite (http://igs-server.cnrs-mrs.fr/~audic/cgi-bin/winflat.pl) können die Signifikanzberechnungen nach dem Test von Audic und Claverie (1997) durchgeführt werden. Jedoch können hierüber nur einzelne Paare abgefragt werden, so daß dieses Vorgehen für ganze Datenreihen kaum in Frage kommt. Des weiteren stellen die Autoren ein Programm für UNIX Rechner zur Verfügung. Diese führt die Berechnungen jedoch auch nur für einzelne Paare durch. Dies reflektiert die Tatsache, daß dieses statistische Vorgehen ursprünglich für EST Sequenzierungsprojekte geplant war, welche wesentlich weniger Transkripte enthielten.

Verwendung in der Literatur

Die Literaturrecherche (Stand 2001) ergibt, daß der Ansatz von Audic und Claverie (1997)

selten verwendet wird (zum Beispiel bei Welle et al. 2000). Margulies und Innis (2000) haben ihn zur Berechnung des *p*-Wertes in der von ihnen entwickelte SAGE-Analyse-Software "eSAGE" übernommen, ohne ihr Vorgehen jedoch näher zu begründen. Allerdings läßt diese Software keine ausschließlich statistische Datenanalyse zu, sondern kann nur für die gesamte Auswertung eines SAGE Projektes inklusive der Schritte vor der statistischen Prüfung verwenden werden. Es wäre folglich erstrebenswert, wenn eine Software entwickelt werden würde, die auf dem Test von Audic und Claverie beruhend auch den rein statistischen Vergleich ganzer SAGE Expressionsprofile gestattet, um im Bereich der Statistik größere Flexibilität zu ermöglichen.

5.2.3.2.3 Vier-Felder-Chi²-Test

Voraussetzungen

Zu den Voraussetzungen des Tests finden sich in der Literatur verschiedene Hinweise. Im allgemeinen wird als Bedingung gefordert, daß die Erwartungswerte pro Zelle mindestens 5 betragen sollten (Bortz 1993[4], S. 159 und Man et al. 2000). Bortz (1993[4], S. 159) weist jedoch darauf hin, daß bei geringeren Erwartungswerten der Test auch dann noch einsetzbar ist, wenn der Umfang der Stichprobe N größer als 7 ist. Sachs (1999[9], S. 451) gibt an, daß sowohl N_1 als auch $N_2 \geq 6$ sein sollten. Die Erwartungswerte pro Zelle sind bei den vorliegenden Daten nur bei 255 Tagpaaren (Daten ohne Sequenzfehlerkorrektur) beziehungsweise bei 391 Paaren (Daten mit Sequenzfehlerkorrektur) größer oder gleich 5[47]. Da jedoch N_1 und N_2 konstant bei 13584 beziehungsweise 13915 bleiben, wird in der vorliegenden Arbeit davon ausgegangen, daß die Voraussetzungen des Testes als erfüllt betrachtet werden können.
Dennoch wäre es for die Anwendung des Vier-Felder-Chi²-Testes günstiger, wenn die Bedingung für die Erwartungswerte beachtet wird. Dies macht den Test besonders geeignet für SAGE Projekte, die eine sehr große Anzahl von Tags sequenziert haben und/oder ein Gewebe oder einen Zustand untersuchen, das/der ein wenig komplexes Expressionsmuster besitzt, so daß auf die Auswertung sehr kleiner Taghäufigkeiten verzichtet werden kann.

[47] Dies entspricht in den vorliegenden Daten exakt den Tagpaaren, auf die das Kriterium "Mittelwert ≥ 5" zutrifft.

Robustheit und Power

Man et al. (2000) untersuchen in der bereits erwähnten Monte-Carlo-Studie zur Robustheit und Power verschiedener statistischer Verfahren, die im Rahmen von SAGE zur Anwendung kommen, auch den Vier-Felder-χ^2-Test. Dieser erweist sich im Vergleich zu Audic und Claveries Test und Fishers Exakt Test als besonders robust. Das heißt, daß das in den Simulationen beobachtete α sich unabhängig vom Expressionsniveau nahe an der festgelegten α-Irrtumswahrscheinlichkeit bewegt, diese jedoch nicht überschreitet. Die ß-Fehlerwahrscheinlichkeit ist im Bereich von Taghäufigkeiten $n \geq 20$ für $N_1 = N_2$ bei allen drei untersuchten Tests sehr ähnlich und niedrig[48]. Im Bereich von $n \leq 15$ besitzt der Chi²-Test einen geringeren ß-Fehler als die anderen beiden Tests. Wenn den Simulationen ungleiche Gesamtmengen und verschiedene Regulationsausmaße zugrunde gelegt werden, weist der Chi²-Test durchgehend den geringsten ß-Fehler auf: Im Bereich kleiner Taghäufigkeiten ($n \leq 15$) liegt er 5 - 10% unter demjenigen des Fishers Exakt Test und des Tests von Audic und Claverie. Bei $n = 15$ liegt er bei 40%, bei $n = 5$ bei 80% und bei $n = 1$ bei 95%. Anzumerken ist, daß die von Man et al. (2000) vorgestellten Ergebnisse nur als Anhaltspunkt dienen und die Werte des ß-Fehlers nicht genau auf die hier vorliegenden Daten übertragen werden können, da diese Werte auf der Grundlage bestimmter Gesamttagzahlen und unterschiedlichster Regulationsmaße entstanden sind.

Die Beobachtungen dieser Monte-Carlo-Studie decken sich mit den explorativen Ergebnissen der vorliegenden Arbeit. Der statistisch abgesicherte Vergleich des Vier-Felder-Chi²-Testes mit dem Test nach Audic und Claverie zeigt, daß die Anwendung des Chi²-Testes statistisch signifikant mehr Tagpaare ergibt, die einen Häufigkeitsunterschied aufweisen, als die Verwendung des Test nach Audic und Claverie. Wenn nur Tagpaare betrachtet werden, die einen Mittelwert von $m \geq 15$ aufweisen, ist kein Unterschied nachweisbar (jeweils 13 statistisch signifikant verschiedene Tagpaare), während die beiden Tests für Paare mit $m < 15$ deutlich unterschiedlich entscheiden (Audic: 54, Chi²-Test: 116).

Welche Implikationen haben diese Ergebnisse für experimentelle Vergleiche zweier SAGE Profile?

Man et al. (2000) empfehlen die Anwendung des Chi²-Testes aufgrund der im Vergleich zu den anderen beiden untersuchten Tests höheren Teststärke und Robustheit. Da der Vier-Felder-Chi²-Test - vor allem in dem Bereich kleiner Taghäufigkeiten - weniger konservativ

[48] Man et al. (2000) geben keine Werte an. Der aus einem Graphen ablesbare ß-Fehler beträgt schätzungsweise 1%.

entscheidet als der Test nach Audic und Claverie, werden bei der Berechnung der statistisch verschiedenen Tagpaare mittels Chi²-Test mehr derartige Paare gefunden werden, als das bei der Verwendung des Tests nach Audic und Claverie der Fall wäre.

Design der Vier-Felder-Tafel

Es gibt Argumente gegen die übliche Aufstellung der Vier-Felder-Tafel (zum Beispiel Man et al. 2000), wie sie auch in der vorliegenden Arbeit erfolgt ist. Audic und Claverie (1997) kritisieren den Mangel an Homogenität, den die Kategorie "Rest" impliziert. Diese besteht aus der Summe der Häufigkeiten all derjeniger Tags eines Expressionsprofils, die nicht das bestimmte Tag sind, das anhand der Tafel untersucht werden soll. Beim Vergleich verschiedener Bibliotheken repräsentiere diese Kategorie unter Umständen unterschiedliche Untergruppen von Transkripten, da die beiden Expressionsprofile verschiedene Gene beinhalten können. Man et al. (2000) wenden dagegen ein, daß ein SAGE Experiment einem Sammelprozeß aus der Population aller Transkripte entspräche und deswegen wahrscheinlichkeitstheoretischer Natur sei. Das bedeute, daß die Tatsache, daß ein Transkript nicht beobachtet wird, nicht beweist, daß es in der Population nicht vorhanden ist. Für die Anwendung einer solchen Vier-Felder-Tafel in der vorliegenden Arbeit trifft das Argument von Audic und Claverie (1997) insofern nicht zu, da die Grundlage der Untersuchung ein einziges Pool ist.

Anwendung

Der Vier-Felder-Chi²-Test ist sehr leicht und schnell mittels eines Tabellenkalkulationsprogrammes auch für große SAGE Projekte durchzuführen.

Verwendung in der Literatur

Die Literaturrecherche ergibt zum gegenwärtigen Zeitpunkt (2001) nur eine Publikation (Michiels et al. 1999), die den hier vorgestellten Test zur Berechnung der statistischen Signifikanzen verwendet.

5.2.3.2.4 Z-Test

Statistische Ansatz

Man et al. (2000) weisen darauf hin, daß für einen Freiheitsgrad $\chi^2 = z^2$ ist. Der Ansatz des Z-Testes nach Kal et al (1999) ist also nicht neu. Er führt zu den selben Entscheidungen wie der Vier-Felder-Chi²-Test. Somit gelten die Aussagen über die Eigenschaften (Robustheit und Teststärke) des Chi²-Tests auch für den Z-Test nach Kal et al. (1999). Kal et al. (1999) weisen nicht auf die Voraussetzungen zur gültigen Anwendung des Tests hin. Sobald die

Taghäufigkeiten unter 6 sinken sind diese jedoch nicht mehr gegeben, so daß sich auch hier Ähnlichkeiten zu den Einschränkungen des Vier-Felder-Chi²-Tests ergeben.

Da jedoch in einem Großteil der SAGE Expressionsprofile Taghäufigkeiten von $n < 6$ oft auftreten, ist die Approximationsformel zur Berechnung von z im Kontext von SAGE vorzuziehen (modifizierter Z-Test). Es ist jedoch auch hier bei der Interpretation der Ergebnisse zu berücksichtigen, daß die Bedingungen des modifizierten Testes für Taghäufigkeiten $n = 0$ nicht erfüllt sind.

Zur Kalkulation des nicht modifizierten Testes wurde ein Windows Programm, SAGEstat, erstellt[49] (Kal et al. 1999). Dieses berechnet für die beobachteten N_1 und N_2 eine Matrix mit kritischen Taghäufigkeiten, die im Sinne eines Konfidenzintervalls (95%, 99% und 99,9%) die oberen und unteren Grenzen für spezifische Taghäufigkeiten angeben. Taghäufigkeiten des zweiten Profils, die außerhalb dieses Intervalls liegen, können auf dem entsprechenden Niveau als statistisch signifikant verschieden betrachtet werden. Wenig Anwender freundlich erscheint hier jedoch, daß die Konfidenzintervalle nicht für alle theoretisch möglichen Taghäufigkeiten berechnet werden, sondern nur etappenweise[50], so daß eine exakte Ermittlung der statistischen Verhältnisse teilweise nicht möglich ist. Auch ist der p-Wert der Tagpaare nicht bekannt. Dieser ist jedoch anhand eines Tabellenkalkulationsprogrammes unabhängig von der Software sehr leicht zu berechnen und für die statistische Entscheidungsfindung zu verwenden, so daß für die statistische Auswertung eines gesamten SAGE Experimentes auf den Einsatz dieser Software verzichtet werden kann. Zumal für kleine Taghäufigkeiten die Verwendung des *modifizierten* Z-Testes der Berechnung des p-Wertes anhand von SAGEstat vorzuziehen ist. Zur Planung von SAGE Experimenten bietet SAGEstat indessen interessante Möglichkeiten. Unter Angabe der zuvor festgelegten α- und ß-Irrtumswahrscheinlichkeit (siehe S. 133) kann ermittelt werden, wie viele Tags pro Profil sequenziert werden müssen, um auf verschiedenen Expressionsniveaus[51] bestimmte Regulationsmaße detektieren zu können. Diese Berechnungen sind auch für den Fall möglich, daß eines der beiden Profile bereits vorhanden ist. Mit SAGEstat steht also eine Software zur Verfügung, die die orientierende Planung effizient durchgeführter SAGE Projekte möglich

[49] SAGEstat wurde der Autorin von J. M. Ruijter (Abteilung für Anatomie und Embryologie der Universität Amsterdam) freundlicherweise zur Verfügung gestellt.

[50] Bis 20 Tags wird das entsprechende Intervall zu jeder Häufigkeit angegeben. Über 20 Tags erfolgen die Angaben in 2er Schritten, über 30 Tags in 3er Schritten und so weiter.

[51] Auch hier sollten aus den diskutierten Test immanenten Bedingungen bezüglich der minimalen gültigen Taghäufigkeiten lediglich Expressionsniveaus abgefragt werden, die diese untere Grenze überschreiten.

macht. Der modifizierte Z-Test ist nach Gleichung 8 einfach mittels eines Tabellenkalkulations-programmes zu berechnen.

Testvergleich

Der modifizierte Z-Test unterscheidet sich deutlich von den anderen drei evaluierten paarweisen Tests. Bei Anwendung dieses Tests werden statistisch signifikant mehr Tagpaare als unterschiedlich erachtet. Wenn jedoch nur diejenigen Tagpaare betrachtet werden, die einen Mittelwert von mindestens fünf aufweisen, verschwindet dieser Unterschied, was den Chi²-Test und den Test nach Audic und Claverie betrifft. Dies ist ein Hinweis darauf, daß das progressive Entscheidungsverhalten des modifizierten Z-Tests sich insbesondere auf den Bereich kleiner Taghäufigkeiten bezieht. Der Unterschied zwischen dem Z-Test und dem Test nach Madden et al. bleibt bestehen, wenn die Resultate des simulierten 5% Niveaus betrachtet werden. Dies spiegelt die Tatsache, daß der Test nach Madden im Gegensatz zum modifizierten Z-Test tendenziell konservativ entscheidet (Kal et al. 1999).

Vergleich der beiden Formen des Z-Testes

Es ist zu sehen, daß die modifizierte Testform sehr viel progressiver entscheidet als die von Kal et al. (1999) dargestellte Variante beziehungsweise der 4-Felder-Chi²-Test. Diese Eigenschaft des auf der Winkeltransformation beruhenden Z-Testes hat zur Folge, daß im Rahmen der Simulation üblicher SAGE Experimente vergleichsweise viele Tagpaare einen statistisch bedeutsamen Unterschied aufweisen (siehe auch Tabelle 23). Ob es sich bei Paaren, bei denen einer der Partner eine Häufigkeit von 0 hat, um eine Verletzung der Voraussetzungen des Z-Testes handelt, die zu falsch positiven Ergebnissen führt, wäre mittels eine Monte-Carlo-Studie zu prüfen (Bortz 1993[4], S. 123ff). Eine solche Studie überschreitet jedoch den Rahmen der vorliegenden Arbeit. Solange die Ergebnisse einer derartigen Studie zur Robustheit des modifizierten Z-Testes nicht bekannt sind, sollte bei der Entscheidung darüber, welche Tagpaare mit einem statistisch nachgewiesenen signifikanten Unterschied in weiteren Untersuchungen betrachtet werden sollen, berücksichtigt werden, daß bei Paaren, bei welchen das Transkript in einem der beiden Profile nicht beobachtet wurde, unter Umständen ein α-Fehler[52] vorliegt. Zusammenfassend ist zu sagen, daß der modifizierte Z-Test exakter und im Kontext von SAGE gültiger anzuwenden ist als die von Kal et al. (1999) vorgestellte Form. Insbesondere bei Projekten, die insgesamt eine große Anzahl an Tags sequenzieren oder - was die Genexpression betrifft - wenig komplexe Gewebe oder Zustände untersuchen, ist der Test zu empfehlen, da auf die Auswertung von Tags mit $n = 0$ verzichtet werden kann.

---

[52] α-Fehler: Die H_1 wird angenommen, obwohl H_0 gilt.

5.2.4 Reliabilität

Der Begriff Reliabilität bezieht sich auf das Ausmaß, in dem Messungen - unter gleichbleibenden Bedingungen durchgeführt - statistisch identisch repliziert werden können.
Im Folgenden wird diskutiert, inwiefern die Reliabilität von SAGE in der existierenden Literatur untersucht wird (5.2.4.1), was das in der vorliegenden Arbeit verwendete Versuchsdesign zur Untersuchung der Reliabilität genau beitragen kann, beziehungsweise ob und wie es die dazu notwendigen formellen Kriterien erfüllt, und auf welche der einleitend vorgestellten Komponenten der Reliabilität es sich dabei bezieht (5.2.4.2). Daran schließen sich die Diskussion der statistschen Ergebnisse dieser Arbeit, insofern sie die Reliabilität von SAGE betreffen (5.2.4.3), und ein Exkurs zur Validierung von SAGE Daten in der Literatur (5.2.4.4) an. Es folgen eine Erörterung derjenigen Aspekte der Praxis von SAGE, die die Reliabilität der Methode beziehungsweise ihr Messung beeinflussen könnten (5.2.4.5), und der Möglichkeiten die Meßgenauigkeit von SAGE zu verbessern (5.2.4.6) inklusive einer abschließenden Evaluation der durchgeführten Sequenzfehlerkorrektur. Es folgt ein Fazit (5.2.4.7) und ein Ausblick (5.2.4.8).

5.2.4.1 Die Reliabilität von SAGE in der Literatur

In der ersten Publikation von Velculescu et al. (1995) findet sich keine Aussagen zur Reliabilität von SAGE. Im Folgenden sollen spätere Veröffentlichungen, die das Thema Reliabilität behandeln, vorgestellt und bewertet werden.
Peters et al. (1999) behaupten, daß SAGE reliabel sei, ohne dies jedoch gesondert zu belegen oder experimentell zu überprüfen. Madden et al. (1997) schließen auf eine zufriedenstellende Reliabilität von SAGE, indem die beobachteten Expressionsstärken einiger "Housekeeping"-Gene und ribosomaler Proteine[53] in den beiden untersuchten Expressionsprofilen vergleichen werden. Diesem Ansatz liegt die weit verbreitete Annahme zugrunde, daß die Expression dieser Gene konstant bleibt - unabhängig von Gewebsart, Entwicklungszustand und pathologischem Hintergrund, obwohl diese These bereits mehrfach hinterfragt (zum Beispiel Velculescu et al. 1999b) und widerlegt worden ist (zum Beispiel Spanakis 1993). Interessant ist in diesem Zusammenhang auch die Veröffentlichung von Michiels et al. (1999) über per SAGE erstellte zerebrale Expressionsprofile, in welcher "Housekeeping"-Gene wie GAPDH und γ-Aktin signifikant reguliert erscheinen. Desssenungeachtet, beziehen sich Madden et al. (1997) auf diese Transkripte und zeigen, daß sich GAPDH (107/92 Tags bei pro Profil circa

[53] Die Daten dieser Transkriptgruppe werden nicht konkretisiert.

30.000 sequenzierte Tags), EF1 (327/396) und exogenes p53 (46/32) nach der von dieser Arbeitsgruppe für SAGE entwickelten Statistik auf dem 5% Niveau nicht statistisch signifikant unterscheiden. Hieraus wird geschlossen, daß SAGE ein reliables Verfahren sei. Wie gezeigt wurde, wäre es angemessen eine solche Fragestellung mit einem Äquivalenztest zu prüfen. In dem vorliegenden Fall ist genaugenommen keine positive Aussage zur Gleichheit der Transkripte möglich. Zudem ist die Teststatistik nach Madden et al. (1997) als konservativ entscheidend bekannt (Kal et al. 1999), was die Nichtannahme von H_1[54] unterstützt. Die erneute Berechnung der Signifikanzen der drei von Madden et al. verwendeten Tagpaare mittels SAGEstat ergibt für EF1 einen *p*-Wert von 0,009, für GAPDH von 0,295 und für p53 von 0,115. Dies heißt, daß eines der Tagpaare (EF1) einen statistisch signifikanten Unterschied aufweist, so daß die globale H_0 nicht angenommen werden kann[55]. Die Werte dieser Publikation lassen demzufolge *nicht* den Schluß zu, daß SAGE reliabel mißt. Zu kritisieren ist außerdem, daß lediglich drei Gene von pro untersuchtem Profil mehr als 9000 verschiedenen Genen verglichen werden. Die Autoren beziehen sich in ihrer Beweisführung also nur auf 0,03% der beiden von ihnen erstellten Expressionsprofile. Zudem handelt es sich hierbei um relativ häufig auftretende Transkripte, so daß die Aussage zur Reliabilität nur auf diese Expressionsniveaus beschränkt hätte getroffen werden können.

Eine Überprüfung der Reliabilität beziehungsweise ihrer Teilaspekte über das gesamte erstellte Expressionsprofil hinweg wäre folglich sinnvoller, wodurch sich auch ein fragliches Hilfskonstrukt wie die Verwendung von "Housekeeping"-Genen umgehen läßt. Von einem derartigen Experiment berichten Bertelsen und Velculescu (1998) in einem Review zu SAGE auf persönliche Kommunikation mit W. Zhou verweisend. Aus ein und derselben RNS Präparation wurden hier verschiedene SAGE Bibliotheken erstellt und miteinander verglichen. Sobald eine genügend große Anzahl an Tags sequenziert worden war, wurden Differenzen, welche zuvor bei geringerer Sequenzierungsmenge beobachtet worden waren, hinfällig, so daß sie auf Stichprobenvariationen zurückgeführt wurden. Die Autoren halten so erwiesen, daß die Messungen von SAGE eine gute Reliabilität aufweisen. Leider wird dieses Experiment nicht mit Zahlen belegt, so daß es nicht nachvollzogen werden kann. Es verweist jedoch auf die beiden Ebenen der Meßgenauigkeit von SAGE: die stochastischen Schwankungen, die mit dem Messen einer Stichprobe einhergehen, und die praktisch bedingte (methodische) Ungenauigkeit, die sich ebenfalls mit der Erhöhung der sequenzierten Tagmenge verringert (siehe S. 166f).

---

[54] H_1: Alternativhypothese, daß die Häufigkeitsverteilungen beider Profile verschieden sind.

[55] H_0 muß bereits erworfen werden, wenn nur ein Tagpaar statistisch signifikant verschieden ist.

Angelastro et al. (2000a) verwenden zum Reliabilitätsnachweis ebenso wie Madden et al. (1997) "Housekeeping"-Gene (EF-1 Varianten) - ohne dies jedoch statistisch zu belegen. Die von der Autorin der vorliegenden Arbeit aus den angegebenen Expressionswerten ermittelten entsprechenden *p*-Werte lauten: EF-1 α 0,336, EF-1 ß 0,699, EF-1 γ 0,724 und EF-1 δ 0,842.[56] Die Verschiedenheit der vier Transkripte in beiden Profilen kann folglich nicht angenommen werden. Da jedoch nicht das gesamte Profil geprüft wird, kann keine allgemeine Aussage zur Reliabilität von SAGE getroffen werden. Angelastro et al. (2000a) verweisen zusätzlich darauf, daß ihre Ergebnisse mit denjenigen aus der Literatur übereinstimmen. Es handelt sich dabei jedoch nicht - wie in der Publikation angenommen - um einen direkten Nachweis der Reliabilität, sondern um einen Hinweis auf die Validität von SAGE.
Die Literatur zur Reliabilität von SAGE gibt also ein widersprüchliches und unvollständiges Bild wider. Diese Lücken versucht die vorliegende Arbeit zu schließen. Inwiefern dies gelungen ist und mit welchem Ergebnis, soll in den folgenden Abschnitten diskutiert werden.

5.2.4.2 Formale Grundvoraussetzungen der Reliabilitätsuntersuchung

Reliabilitätsmessung allgemein

Reliabilität, wie sie einleitend definiert wurde, erfordert die statistische Gleichheit der Resultate einer mehrfach angewandten Methode. Die Prüfung dieser Gleichheit sollte unter den identischen Verhältnissen am selben Material geschehen, so daß sämtliche Schwankungen der Daten auf Anwender oder Methode zurückgeführt werden können. Diesem Ideal gerecht zu werden, ist jedoch in der Praxis von SAGE nur schwer möglich. Im Folgenden soll gezeigt werden, welche Probleme im Rahmen der vorliegenden Arbeit bei der Umsetzung des theoretischen Konzeptes der Ermittlung von Reliabilität entstanden.
Hier noch mal der Versuchsaufbau: Es wurde aus vier Mäusegroßhirnen von gesunden männlichen Tieren derselben Rasse, Alters,- und Gewichtsklasse die Gesamt-RNS extrahiert und vereinigt. Diese Transkriptgrundpopulation wurde zweigeteilt und parallel per SAGE untersucht, wobei pro Gruppe Zufallsstichproben von mehr als je 15000 Tags sequenziert wurden.
Da bei dem hier geschilderten Versuchsaufbau der zweite Durchlauf parallel erfolgte, kann die Bedingung "gleiche Verhältnisse" als weitestgehend erfüllt erachtet werden. Minimale Schwankungen können jedoch nie gänzlich ausgeschlossen werden. Das gilt insbesondere für

---

[56] Angelastro et al. (2000a) geben keine entsprechenden *p*-Werte an. Die hier genannten Werte wurden per SAGEstat errechnet. Das α_{KORR} nach Bonferroni beträgt 1,25% (5% / 4).

langwierige Vorgänge wie dem zweiten NlaIII Verdau (siehe S. 65) oder der Ligation der Tags zu Konkatemeren (siehe S. 66), welche nicht an einem Tag zeitnah erfolgen konnten. Das bedeutet, daß in der Praxis in diesen Punkten bezüglich des Aspektes "gleiche Verhältnisse" ein gewisser Unsicherheitsfaktor besteht, der nicht quantifiziert werden kann.

Schwieriger gestaltet es sich mit der zweiten Forderung: Durchführung der Messungen am *selben* Material. Da im Zuge eines SAGE Durchlaufes die orginale RNS aufgebraucht wird, läßt sich SAGE - wie alle anderen Methoden zur Messung der Genexpression auch - nicht mehrfach am selben Material durchführen. Um sich dieser Bedingung der Reliabilitätsmessung anzunähern, wurde ein Pool hergestellt, das die RNS mehrerer Individuen in sich vereinigt. Auf diese Weise wurde eine sehr große Transkriptgrundpopulation geschaffen. Es wurde angenommen, daß durch die enorme Anzahl der darin enthaltenen RNS-Moleküle selbst gering exprimierte Gene dieselbe Chance haben, bei der Teilung des Pools in den beiden parallelen Durchläufen präsent zu sein.[57]

Die dritte Bedingung der Reliabilitätsmessung fordert die mehrfache Anwendung der zu evaluierenden Methode. Die Reliabilität einer Methode ist durch das Ausmaß des zufälligen Meßfehlers charakterisiert, das sich anhand von Meßwiederholungen bestimmen läßt.[58] Je stärker dieser Fehler streut, um so größer muß die Anzahl der Beobachtungen sein, bis sich die Schätzung des Parameters stabilisiert. Diese Stabilisierung läßt sich auf unterschiedlichen Wegen erreichen. Im Rahmen der 'Framingham Heart' Studie (Dawber 1980) wurden zum Beispiel Reliabilitätsstudien zur Registrierung einiger Laborwerte durchgeführt. Diese kamen zu dem Schluß, daß beispielsweise eine Stichprobengröße von $N = 10$ bei jeweils zehn Wiederholungen pro Individuum genauso effizient den Meßfehler bestimmen kann wie nur je zwei Messungen von 100 Individuen. Auf SAGE übertragen, hieße das entweder viele (zum Beispiel 100) RNS-Pools zu teilen und jeweils die beiden Hälften parallel per SAGE zu messen oder wenige (zum Beispiel 10) Pools mehrfach zu teilen (10 mal) und ebenfalls parallel zu behandeln. Dies macht deutlich, daß dem vorgestellten Versuchsaufbau eine Dimension fehlt, da nur eine einzige Stichprobe ($N = 1$) zweimal gemessen wurde. Die genannten Versuchsdesigns würden jedoch den Rahmen der Möglichkeiten des hier vorgestellten Projektes sprengen. Diese Einschränkung hat zur Folge, daß die sinnvolle

---

[57] Zufällige Schwankungen sind jedoch nicht ganz auszuschließen. Diese sind jedoch beispielsweise auch vorhanden, wenn dasselbe Individuum zur Reliabilitätsprüfung eines Tests über die Zeit wiederholt untersucht wird.

[58] Der Anteil des wahren Wertes bleibt bei der Meßwiederholung konstant, während der Anteils des zufälligen Fehlers variiert und so über die Gesamtvarianz der Meßergebnisse ermittelt werden kann (Lienert 1994, S. 176).

Anwendung standardmäßiger statistischer Kennwerte der Reliabilität wie beispielsweise der Fehlervarianz, dem Retest-Reliabilitätskoeffizienten oder "Intraclass Correlation" Koeffizient in der vorliegenden Arbeit nicht erfolgen kann, da diese mehr Stichproben erfordern.
Das bedeutet, daß in der vorliegenden Arbeit der Nachweis der Reliabilität nur *indirekt* erfolgen kann, indem die beiden ermittelten Merkmalsvektoren K1 und K2 daraufhin untersucht werden, ob ihre Verteilungen homogen sind (globale H_o), was - im dem Fall, daß H_1 nicht angenommen werden kann - ein Hinweis darauf wäre, daß die Schritte eines SAGE-Durchlaufes keine Ungenauigkeit eingeführt haben. Die gleiche Vorgehensweise findet sich bei Spinella et al. (1999), um die Reliabilität der von ihnen entwickelten (SAGE ähnlichen) Methode zur Messung von Genexpression zu prüfen. Es werden ebenfalls zwei Expressionsprofile aus einem gemeinsamen RNS-Pool erstellt und die Tags mittels Vier-Felder-Chi²-Test paarweise auf statistische Unterschiede untersucht.[59]

Die Messung der verschiedenen Aspekte der Reliabilität

Die verschiedenen Aspekte von Reliabilität sind Reproduzierbarkeit, Wiederholbarkeit und Stabilität. Im Folgenden wird dargestellt, inwiefern diese vom Design der vorliegenden Studie erfaßt werden.

a) Stabilität

Beim Begriff der Stabilität steht die biologische Variabilität eines Individuums über die Zeit im Vordergrund. Da sämtliche molekular-biologischen Methoden zur Messung der Genexpression im Messen den Gegenstand ihrer Untersuchung zerstören, kann das Ausmaß dieser *intra*individuellen Variabilität nie ermessen werden. Theoretische Überlegungen, Computersimulationen, chemische Modelle und indirekte Untersuchungen zeigen folgendes: Die natürlichen Schwankungen, die die biologische Variabilität bedingen, haben im Kontext der Genexpression - je nach Stärke des Promotors (Kierzek 2000) - stochastische und pulsatile Eigenschaften, was bedeutet, daß die Transkription großen Fluktuationen unterworfen sein kann (Newlands 1998, Arkin 1998). Diese Schwankungen können sich je nach Gen (Spanakis 1994), nach Tageszeitpunkt (Lavery 1997), nach Zelltyp und Entwicklungsphase unterschiedlich gestalten. Selbst ständig vorhandene Proteine können von nicht-kontinuierlich aktiven Genen kodiert werden (Kierzek 2000, Newlands 1998). Dies hat zur Folge, daß die

[59] Es wurden allerdings nur Tags verglichen, die mindestens eine Häufigkeit hatten, die 0,01% der Gesamtanzahl entsprach. Auf die Daten der vorliegenden Arbeit übertragen, hieße dies, sämtliche Singletons von der Reliabilitätsuntersuchung auszuschließen.

zeitlichen Muster einer spezifischen Expression in einzelnen Zellen eines Zellverbandes sich unter Umständen wenig gleichen und unberechenbar sein können (McAdams 1997).
Diese natürliche Variabilität beeinflußt vor allem dann sämtliche Aspekte der Reliabilität, wenn Messungen an *einzelnen Individuen* durchgeführt werden, da sie Teil der *inter*individuellen Variabilität ist. Um diese Einflußgröße statistisch in den Griff zu bekommen, wäre es notwendig, die Werte mehrerer Individuen statistisch zu mitteln (Spanakis 1994). Da jedoch ein Teil der Anwendungen von SAGE auf die Verwendung immer kleinerer Mengen zielt (Datson et al. 1999, Peters et al. 1999, Ye und Zhang 2000) - bis hin zum Vergleich von Einzelzellanalysen (Brady 2000), ergibt sich das Problem, daß Einzelfälle verglichen werden, deren intra- und interindividuellen Variabilitäten unbekannt sind - ein statistisch fragliches Unterfangen. Wenn die RNS mehrerer Individuen vereinigt wird, um eine ausreichend große Menge für einen SAGE-Durchlauf zur Verfügung zu haben, können interindividuelle Schwankungen eventuell ausgeglichen werden. Es besteht jedoch auch die Gefahr, daß einzelne stark abweichende Expressionsmuster die Resultate verzerren (Welle et al. 2000). Auch hier ist das Problem, daß die Variabilitäten nicht bekannt sind und so nicht beim Vergleich von Expressionsprofilen bedacht werden können.
In der vorliegenden Arbeit wurden die beiden Hälften eines einzigen RNS-Pools verglichen, so daß der Faktor Stabilität keine Rolle spielt, deswegen aber auch nicht erfaßt werden kann.

b) Reproduzierbarkeit

Die Höhe der Reproduzierbarkeit einer Methode ist unmittelbar abhängig von der Exaktheit der Arbeitsweise des *Untersuchers*. Da im Laufe der Durchführung von SAGE verschiedene Untersucher beteiligt waren, ist es nicht möglich die Reproduzierbarkeit, die sich per definitionem auf die Gleichheit der Resultate wiederholter Durchläufe eines einzigen Untersuchers bezieht, einzuschätzen. Es kann nicht im Sinne effizienter Forschung und sinnvollem Zeit- und Arbeitskraftmanagement sein, eine derart arbeitsintensive Methode wie SAGE von einer einzigen Person durchführen zu lassen, so daß die Überprüfung der Reproduzierbarkeit von SAGE nicht praktikabel zu sein scheint. Durch standardisiertes Arbeiten soll erreicht werden, daß Schwankungen der Reproduzierbarkeit möglichst gering ausfallen.

c) Wiederholbarkeit

Das Ausmaß der Wiederholbarkeit wird per definitionem nur von der Meßmethode selbst beeinflußt. Die Abschätzung der Größe dieses Einflusses, erfordert die Prüfung der Gleichheit

der Resultate, welche ein Verfahren liefert, indem dieses unter gleichen Bedingungen (das hießt auch von einer Arbeitsgruppe) am selben Material erneut durchgeführt wird. Dies entspricht dem Versuchsaufbau der vorliegenden Arbeit.

5.2.4.3 Die statistische Überprüfung der Reliabilität

Die statistische Prüfung der Reliabilität von SAGE erfolgte in der vorliegenden Arbeit indirekt. Die beiden Profile wurden auf Homogenität (Chi-Test für k x 2 Felder-Tafeln mit Monte-Carlo-Simulationen) geprüft. Dabei wurde davon ausgegangen, daß die beiden Expressionsprofile, die parallel erstellt worden waren, die statistisch gleichen Resultate liefern würden, wenn die Meßgenauigkeit von SAGE hinreichend groß wäre. Ein solches Vorgehen kann keine exakten quantitativen Aussagen zur Reliabilität liefern, sondern lediglich qualitative Hinweise geben. In dem Fall, daß H_1 (statistische Ungleichheit) angenommen werden kann, wäre SAGE als nicht reliabel einzustufen.

Dies ist in der hier vorliegenden Untersuchung der Fall. H_1 mußte angenommen werden, da sich die beiden Profile - insgesamt betrachtet - als statistisch nicht identisch erwiesen. Das bedeutet jedoch nicht zwangsläufig, daß die Reliabilität von SAGE schlecht ist. Dies soll im Folgenden begründet werden.

Da SAGE ein "sammelndes" Verfahren ist, das aus der Grundpopulation eines Transkriptoms Stichproben entnimmt, sind in Abhängigkeit von der Größe dieser Stichprobe gewisse stochastische Schwankungen vorhanden (siehe dazu auch S. 22). Im Fall der vorliegenden Arbeit könnten die Abweichungen der Transkriptverteilungen von K1 und K2 entweder auf eine tatsächlich vorhandene Ungenauigkeit verweisen, die im Laufe eines SAGE Durchlaufes entsteht (für eine genaue Analyse der molekularbiologischen Praxis siehe S. 166f), oder durch die Stichprobenvariabilität entstanden sein. Diese ist im hier diskutierten Projekt aufgrund der Komplexität des untersuchten Gewebes und der relativ kleinen Stichprobe vermutlich hoch. Dies bedeutet, daß in der vorliegenden Arbeit keine endgültige Aussage zur Reliabilität möglich ist - bis auf die folgende: Im vorliegenden Kontext (geringe Stichprobe, komplexes Gewebe) liefert SAGE keine statistisch identischen Tagzahlen und kann nicht als reliabel messend eingestuft werden. Ein weiterer Gesichtspunkt ist, daß ein Chi²-Test (mit Variante: Kontingenzkoeffizient) verwendet wurde, welcher mit wachsendem Stichprobenumfang[60] schnell signifikant wird (mündliche Mitteilung K. Wernecke), so daß progressiv zugunsten der Alternativhypothese entschieden wird.

---

[60] Dies bezeiht sich auf die Gesamttagzahl.

Diskussion der Resultate des Kontingenzkoeffizienten

Die statistische Prüfung der Reliabilität von SAGE wurde in der vorliegenden Arbeit durch die Analyse der Daten mit einem zweiten statistischen Modell ergänzt. Zusätzlich zur Prüfung der beiden Profile auf Homogenität (siehe vorherigen Abschnitt) wurde untersucht, ob zwischen ihnen ein Zusammenhang besteht (Kontingenzkoeffizient).

Diese ergänzende Betrachtung der beiden Profile als verschiedene Merkmalsausprägungen einer Stichprobe anhand einer Kontingenztafel und der Berechnung eines Zusammenhangmaßes (Cramers V), führt zur Bestätigung der Hypothese, daß die beiden Gruppen voneinander stochastisch abhängig sind. Zu berücksichtigen ist dabei, daß möglicherweise eine Inhomogenitätskorrelation bestehen könnte. Aufgrund weit auseinanderliegender Werte entsteht hierbei ein künstlicher Korrelationseffekt, der bei Betrachtung nahe beieinanderleigender Wertegruppen verschwindet.

Im Folgenden werden die Resultate der Untersuchung des Zusammenhangsmaßes Cramers V genauer dargestellt. Nach Bortz (1990, S. 60) ist die über einen Kontingenzkoeffizienten definierte Reliabilität einer biologischen Beobachtung hoch, wenn sie 0,9 erreicht. Bei 0,7 ist sie zufriedenstellend und bei 0,5 ausreichend.[61] Wie sind in diesem Bewertungsschema die Ergebnisse der vorliegenden Arbeit unter der hypothetischen Annahme, daß keine Inhomogenitätskorrelation vorliegt, einzustufen?

Wenn die gesamten Datensätze betrachtet werden, kann die Reliabilität von SAGE - wie sie hier gemessen und definiert wurde - als zufriedenstellend bis ausreichend bezeichnet werden. Dies gilt sowohl für die Daten ohne Sequenzfehlerkorrektur als auch für diejenigen mit Korrektur, wobei letztere dem Augenschein nach einen etwas niedrigeren Kontingenzkoeffizienten aufweisen (0,681 versus 0,637).[62] Werden die Datensätze anhand der Größe des Mittelwertes aufgeteilt (m < beziehungsweise ≥ 5) ist zu sehen, daß die Reliabilität in beiden Datensätzen für die Tagpaare, die einen Mittelwert von größer oder gleich fünf haben, weiterhin als zufriedenstellend bis ausreichend eingestuft werden kann (0,693 und 0,642).[63] Tagpaare dagegen, deren Mittelwert kleiner als fünf ist, können nicht als

---

[61] In psychologischen Tests dagegen wird mindestens eine Reliabilität von 0,8 gefordert (Bortz und Döring 1995², S. 184) Der Bereich von 0,8 bis 0,9 gilt als mittelmäßig.

[62] Es müßte ein spezielles Verfahren für den statistischen Vergleich der hier ermittelten Kennwerte entwickelt werden. Dies kann im Rahmen der vorliegenden Arbeit nicht geleistet werden, so daß die Werte einander nur gegenübergestellt werden können ohne zu prüfen, ob sie statistisch signifikant verschieden sind.

[63] Die Ergebnisse der Prüfung der Gesamtverteilungen für Tagpaare mit m ≥ 5 beziehungsweise 10 auf Homogenität mittels k x 2- Felder-Chi²-Test mit Computersimulationen zeigt jedoch weiterhin, daß die beiden

reliabel gemessen eingestuft werden (0,316 und 0,286). Es scheint also eine Abhängigkeit der Reliabilität von SAGE von der beobachteten Häufigkeit der Tagpaare vorzuliegen. Dieses Resultat kann verschiedene Ursachen haben. a) Die Streuung (Fehlervarianz) ist im Bereich kleiner Häufigkeiten tatsächlich größer. Darüber kann jedoch im Rahmen der vorliegenden Arbeit keine zuverlässige Aussage gemacht werden, da dazu mehrere Profile in die Berechnungen einbezogen werden müßten. b) Es liegt doch eine Inhomogenitätskorrelation vor. Das heißt, daß bei Betrachtung der *Gesamt*profile beziehungsweise der Tags mit $m \geq 5$ eine so starke Inhomogenität der Werte vorliegt (0 beziehungsweise 5 bis über 300), daß dadurch der Kontingenzkoeffizient künstlich hoch wird. Während die Daten bei der Berechnung von Cramers V für Tags, deren Mittelwert $m < 5$ ist, homogener sind (0 - 4), so daß in diesem Fall kein Korrelationseffekt durch Lageunterschiede erzielt wird.[64]

Diskussion des Zusammenhangs von Reliabilität und Taghäufigkeit

Die Frage, ob sich die Größe der Fehlerschwankungen mit der Höhe des Expressionsniveaus ändert, kann in der vorliegenden Arbeit nicht klar beantwortet werden, da aufgrund des Versuchsdesigns (indirekte Reliabilitätsmessung) nur qualitative Aussagen gemacht werden können (siehe S. 158). Wenn die molekular-biologischen Praxis von SAGE betrachtet wird (siehe unten), lassen sich keine theoretischen Erklärungen für einen solchen Zusammenhang finden. Da jedoch größere Häufigkeitswerte - bei angenommener konstanter Fehlervarianz - Häufigkeitsschwankungen besser tolerieren als kleine Werte und das Problem der Stichprobenvariabilität bei Transkripten niedriger Expressionsniveaus eine größere Rolle spielt, erscheint in praxi die Betrachtung von Transkripten, welche häufig auftreten, genauer. Eine absolute Schwankung von beispielsweise ± 4 Tags führt zu starken Meßunsicherheiten, wenn der wahre Wert eine Transkriptes bei 5 Tags liegt, während ein Tag, das realiter 100 mal vorhanden ist, auch durch einen Wert von 97 oder 104 hinreichend genau erfaßt zu werden könnte. Dieser Zusammenhang führt zu der Empfehlung, nur Tagpaare, die eine gewisse Häufigkeit aufweisen, in die Auswertung von SAGE Projekten einzubeziehen. Dies entspricht dem bereits diskutierten Kriterium der "minimalen Taghäufigkeiten". Um bei obigem Beispiel zu bleiben: Bei Schwankungen von ± 4 kann ein Tagpaar mit den Werten 1/9

Gruppen K1 und K2 unterschiedlich sind.

[64] Um den Effekt einer Inhomogenitätskorrelation zu vermeiden, müßten mehrere SAGE Profile erstellt werden, so daß für jedes Tag eine Meßreihe entsteht. Auf diese Weise könnte das Zusammenhangsmaß für jedes Tagpaar getrennt berechnet werden.

unreguliert sein (wahrer Wert wäre in beiden Fällen zum Beispiel 5) oder aber auch wirklich reguliert. Wären in diesem Beispiel Mengen der Größenordnung der vorliegenden Arbeit sequenziert worden, wäre das Tagpaar statistisch signifikant verschieden. Es entspräche auch dem Kriterium "größer zweifach reguliert". Dies zeigt, wie wichtig es ist, zusätzlich zur statistischen Signifikanz, dem Kriterium einer Mindestregulation auch das Kriterium "minmale Taghäufigkeit" zu beachten. Im Fall sehr kleiner Taghäufigkeiten entsteht sonst ein Bereich, in dem häufig falsch positive Resultate zu finden sind.
Finden sich zu diesem Aspekt von SAGE Hinweise in der Literatur?
Vingron und Hoheisel (1999) weisen in einem Review zu SAGE darauf hin, daß reliable Schätzungen der Taghäufigkeiten erst ab einer bestimmten Menge an insgesamt sequenzierten Tags stattfinden können. Audic und Claverie (1997) behaupten, daß reliablere Aussagen über Tags getroffen werden können, deren absolute Häufigkeiten höher sind. Beiden Aussagen ist gemeinsam, daß sie den oben erörterten Zusammenhang von Schwankungen der Meßwerte und deren Größe mit einer Veränderung der Reliabilität gleichsetzen - ohne dies jedoch zu belegen oder Zahlen zu nennen. Ein weiterer Hinweis zu diesem Thema findet sich bei Ishii et al. (2000). Diese Publikation vergleicht die Resultate von SAGE und einer Chip-Technik. Hier geht es um die Übereinstimmung zweier Methoden. Die Reliabilität beider ist eine der notwendigen Voraussetzungen für ihre Übereinstimmung, weswegen an dieser Stelle ein Ergebnis der Studie erwähnt werden soll. Die Messungen beider Methoden sollen näher beieinander liegen, wenn höhere Expressionsniveaus verglichen werden.[65] Auch hier läßt sich hinsichtlich der Meßgenauigkeit keine Aussage zur Ursache dieser Beobachtung machen.

Zusammenfassend läßt sich sagen:
Die Reliabilität von SAGE kann den Ergebnissen der vorliegenden Arbeit entsprechend - das heißt im vorliegenden Kontext - nicht als gut eingeschätzt werden. Es kann jedoch keine Aussage dazu gemacht werden, ob dies der Methode selbst, das heißt ihrer molekularbiologischen Praxis und der Datenaufbereitung, anzulasten ist oder einer großen Stichprobenvariabilität. Anhand welcher Studiendesigns diese Fragen beantwortet werden könnten, soll im Abschnitt Ausblick zusammengefaßt werden (ab S. 175).

Dieses Ergebnis der vorliegenden Reliabilitätsstudie führt dazu, sich den eingangs postulierten Erwartungen erneut zuzuwenden und zu analysieren, ob deren Grundlagen unter Umständen fehlerhaft sein könnten. Dies soll im folgenden Exkurs zur Validität geschehen.

[65] Das gleiche gilt, wenn die Häufigkeiten stark regulierter Gene miteinander verglichen werden.

5.2.4.4 Exkurs: Validierung von SAGE Daten in der Literatur

Einer der Gründe dafür, zu Beginn der vorliegenden Arbeit die Hypothese aufstellen zu können, daß die Reliabilität von SAGE als gut einzuschätzen ist, war, daß sich die per SAGE erstellten Expressionsdaten in der Literatur gut per Northern Blot oder anderer Methoden validieren ließen. Da das Vorhandensein von Validität an die Existenz eines gewissen Maßes von Reliabilität geknüpft ist (siehe S. 24), wurde letztere bisher als gegeben angenommen.Wenn die Reliabilität von SAGE - wie im vorliegenden Kontext - sich nicht als eindeutig gut darstellt, wie lassen sich dann die guten Ergebnisse der Validierungen in der Literatur erklären?

Eine zweite Durchsicht der Literatur ergibt methodische Mängel der statistischen Prüfungen der Validität von SAGE. Da die diese jedoch nicht das Thema der vorliegenden Arbeit ist, soll an dieser Stelle im Folgenden lediglich beispielhaft auf zwei frühe Veröffentlichungen der Arbeitsgruppe, die SAGE entwickelt hat, eingegangen werden.

In der ersten Publikation zu SAGE (Velculescu et al. 1995) erfolgte die Validierung der Taghäufigkeiten der hier erstmals vorgestellten Methode per Vergleich mit den Hybridisierungsergebnissen von cDNS Bibliotheken. Dies geschah nur für 10 verschiedene Gene (von 428 insgesamt beobachteten) und ohne statistischen Prüfung. Die Tatsache, daß sich die relativen Quantitäten der beiden Methoden per Augenschein ähnlich waren, genügte, um zu der Schlußfolgerung zu kommen, daß eine "gute Übereinstimmung" (Velculescu et al. 1995) zwischen beiden besteht.

1997 geben Velculescu et al. an, daß die Korrelation der Expressionsniveaus von einigen Genen, die sich in einer Untersuchung zur Genexpression der Hefe als stark exprimiert und/oder statistisch signifikant reguliert erweisen haben, mit Daten, die per Northern Blot und PhosporImager erstellt wurden, hoch ist. Das angegebene Bestimmtheitsmaß r^2 liegt bei 0,97. Wenn dieser Kennwert für die beiden Profile (Daten ohne Sequenzfehlerkorrektur) der vorliegenden Arbeit berechnet wird, ergibt sich ein Wert von 0,84 ($r = 0,91$), also ebenfalls eine sehr hohe Korrelation, die sich jedoch bei Anwendung der adäquaten Statistik nicht bestätigt[66]. Der hier verwendete Korrelationskoeffizient (Produkt-Moment-Korrelation) dient der Messung eines linearen Zusammenhangs metrisch skalierter Merkmale[67], wohingegen

---

[66] Ein hoher Wert von r kann beispielsweise allein dadurch zustande kommen, daß die Werte weit gestreut sind (Altman 1991). Dies ist bei Velculescu et al. (1997) der Fall; es finden sich Taghäufigkeiten von 0 bis 561.

[67] Skalentypen: Nominalskala (Unterteilung nach Gruppen, Beispiel: Tags bei SAGE, Krankheitsklassifikationen), Ordinalskala (Rangdaten, Beispiel: Schulnoten), Intervallskala (metrische

adäquate Zusammenhangsmaße für nominal skalierte Häufigkeitsdaten, wie sie im Falle von SAGE vorliegen, Kennwerte wie beispielsweise die verschiedenen Kontingenzkoeffizienten wären (Bortz 1993, S. 215f).[68] Die Richtigkeit der von Velculescu et al. (1997) getroffenen Aussage zur Validität kann also nicht beurteilt werden. Außerdem handelt es sich bei beiden Untersuchungen um keine *vollständige* Überprüfung der Validität von SAGE, sondern nur um eine Teilbetrachtung "repräsentativer" (Velculescu et al. 1997) Gene, nämlich stark exprimierter und statistisch signifikant regulierter. Eine Untersuchung der Validität gesamter SAGE Profile, die sehr aufwendig wäre, in Kombination mit derjenigen Statistik, die dem Skalenniveau der Daten und dem Untersuchungsziel angemessen ist, steht noch aus.
Diese kurze Diskussion sollte andeutungsweise begründen, weshalb zwischen den veröffentlichten Resultaten zur Validität von SAGE und den hier dargestellten Ergebnissen zur Reliabilität nur ein scheinbarer Widerspruch besteht.

5.2.4.5 Inhaltliche Bewertung der Reliabilität von SAGE

Es folgt eine Zusammenfassung methodischer Probleme von SAGE unter dem Gesichtspunkt ihrer Auswirkung auf die Reliabilität des Verfahrens.
Zur Erinnerung: Meßungenauigkeit im Sinne der methodischen Reliabilität entsteht durch zufällige Fehler. Derartige Fehlermöglichkeiten in der Praxis von SAGE sollen im folgenden dargestellt werden. Dabei ist folgendes zu beachten: "Es liegt in der Natur der zufälligen Fehler, daß über ihre Ursachen nur wenig ausgesagt werden kann. " (Hart et al. 1997[7], S. 62).
Ein SAGE-Durchlauf kann in drei Phasen eingeteilt werden:

- von der Extraktion der Gesamt-RNS bis zur Fertigstellung sequenzierbarer Konkatemere,
- die Sequenzierung,
- die Auswertung mit a) Erstellung der endgültigen Tagliste, b) der Homologierecherche und c) der statistischen Auswertung.

Da diesen Phasen sehr unterschiedliche Prozesse zugrunde liegen, ist es sinnvoll, bei der inhaltlichen Betrachtung der Reliabilität von SAGE diese getrennt zu bewerten.

Meßwerte, Beispiel: Temperatur in °C), Verhältnisskala (metrische Meßwerte mit absolutem Nullpunkt, Beispiel:Temperatur in Kelvin) (Bortz 1993, 24ff).

[68] Zur Vorgehensweise bei Validierungen vergleiche zum Beispiel Altmann (1991). Dort wird darauf hingewiesen, daß es sich hierbei statistisch *nicht* um eine Korrelation handelt.

Erste Phase: Von der Extraktion der Gesamt-RNS bis zur Fertigstellung der Polytags

Generell sind molekular-biologische Technologien durch Schwankungen der Umgebungstemperatur, der Enzymaktivitäten (besonders von NlaIII), der Pipettiervolumina et cetera für zufällige Fehler anfällig. Dies spielt gerade bei einer aufwendigen Methode wie SAGE, die viele komplexe Schritte beinhaltet, eine gewichtige Rolle und trägt dazu bei, daß deren Meßgenauigkeit sinken kann. Bei der Diskussion der Etablierung von SAGE wurde außerdem deutlich, daß es über diese allgemeinen Punkte hinaus besonders vulnerable Schritte gibt, die die Wiederholbarkeit beeinflussen können.

Zweite Phase: Sequenzierung

Der Sequenzfehler hat einen großen Einfluß auf die Meßgenauigkeit von SAGE, da sich durch diesen zufälligen Fehler die ermittelten Häufigkeiten direkt verändern. Die für die vorliegenden Daten aus den Linkersequenzierungen geschätzte *maximale* Wahrscheinlichkeit für mindestens einen Fehler liegt bei 31%, was die Reliabilität von SAGE deutlich einschränkt. Hier ist noch mal die Wichtigkeit einer Sequenzierung in beiden Richtungen zu betonen[69]. Die entstehenden Mehrkosten senken aufgrund der präziseren Ergebnisse den Kostenaufwand der Folgestudien.

Dritte Phase: Auswertung

Um aus einem SAGE Projekt sinnvolle Aussagen ableiten zu können, ist eine Auswertung der SAGE Rohdaten in drei Schritten - Erstellung der endgültigen Tagliste, Homologierecherche und Statistik - notwendig. Auch hier haben bestimmte Schritte oder Konstellationen besondere Relevanz für die Reliabilität von SAGE. Im folgenden sollen die ersten beiden Schritte beleuchtet werden.

a) Erstellung der endgültigen Tagliste

Hierbei ist die Elimination der Linkerartefakte hervorzuheben. Durch die Entfernung der Linkerderivate wie es in der vorliegenden Arbeit geschehen ist, ergibt sich die Situation, daß ein Linkertag, wenn es sich aufgrund von mehrfachen Sequenzlesefehlern um mehr als eine Base von den Orginallinkern unterscheidet, im Datensatz verbleibt. Die geschätzte Wahrscheinlichkeit für das Vorhandensein derartiger Derivate beträgt für die vorliegenden Daten 5% (siehe S. 92). Wenn ein solches Derivat zufälligerweise mit einem bereits

[69] Zu den Ergebnissen der Sequenzfehlerkorrektur siehe S.164.

vorhandenen Tag übereinstimmt, verfälscht es dessen Häufigkeitswert, so daß die Meßgenauigkeit negativ beeinflußt wird. Wenn ein solches Linkerartefakt mit mehrfachen Sequenzfehlern dagegen mit keinem Tag aus einer RNS Population korrespondiert - was allerdings erst auffallen könnte, wenn sämtliche mögliche Tags eines Gewebes oder einer Spezies bekannt sind - kann es ebenfalls die Reliabilität von SAGE senken, da es sein Zustandekommen einem zufälligen Fehler verdankt, der in den beiden Profilen unterschiedlich ausfallen kann. Über Tags, die realiter von einer Boten-RNS abstammen, die jedoch im Rahmen der Entfernung von Linkerartefakten eliminiert werden, weil sie per Zufall von einer Base abgesehen mit den Linkersequenzen übereinstimmen, kann bezüglich der Reliabilität keine Aussage getroffen werden, da sie nicht gemessen werden.[70]

b) Homologierecherche

Die Reliabilität, wie sie in der vorliegenden Arbeit betrachtet wird, bezieht sich auf die Messung der Taghäufigkeiten. Die Reliabilität der Homologierecherche selbst wird statistisch nicht erfaßt. An dieser Stelle soll auf Aspekte eingegangen werden, die die Genauigkeit der Homologierecherche und auch das quantitative Resultat von SAGE beeinflussen können.

Wenn ein Tag zum Beispiel aufgrund eines unvollständigen NlaIII Verdaus als inneres Tag (5% bei Welle et al. 1999) erscheint, jedoch als solches bei der Auswertung nicht erkannt und demzufolge falsch zugeordnet wird, wird die Genauigkeit von SAGE vermindert, da die Häufigkeit des eigentlichen Transkriptes zu gering ausfällt, während diejenige eines anderen Tags fälschlicherweise ansteigt oder aber ein neues Tag entsteht. Durch die Zuordnung der Tags zu Genen kann die induzierte Meßungenauigkeit noch ausgeweitet werden, wenn ein neu entstandenes Tag einem bislang reliabel bestimmten Gen durch die Homologierecherche mit zugeordnet wird und so dessen Häufigkeitswert verzerrt.

Ein zweiter Aspekt, der im Kontext von üblichen SAGE Experimenten relevant sein kann: Wenn Expressionsprofile verglichen werden, welchen - im Gegensatz zur vorliegenden Arbeit - jeweils unterschiedliche Individuen zugrunde liegen und in deren Transkriptomen verschiedene Polymorphismen (SNPs, Spleißvarianten, siehe S. 98ff) vorliegen, dann entstehen für ein und dasselbe Gen (teilweise) unterschiedliche Tags. Hinsichtlich der Reliabilität der Messung sind dann drei Konstellationen möglich: Werden die Tags der beiden Expressionsprofile unmittelbar verglichen, ergibt sich eine erhöhte Meß*un*genauigkeit. Dies hat zur Folge, daß der Vergleich von Tagprofilen einzelner Individuen nicht reliabel sein kann und deswegen nicht durchgeführt werden sollte. Werden die Genprofile nach erfolgter

[70] Hier liegt ein systematischer Fehler vor, der die Validität von SAGE negativ beeinflußt.

Homologierecherche verglichen, können - wenn die Polymorphismen in der Datenbank enthalten sind - die Verhältnisse wieder ausgeglichen sein. Wenn jedoch nicht alle Polymorphismen in die UniGene Cluster aufgenommen sind - wovon zum gegenwärtigen Zeitpunkt ausgegangen werden kann - oder aber durch Polymorphismen eine NlaIII Erkennungssequenz zerstört oder eine neue einführt wurde, werden Tags gar nicht oder falsch zugeordnet. Dann ergibt sich auch beim Vergleich der zugeordneten Tags ein quantitatives Reliabilitätsproblem. Dies macht deutlich, daß in der vorliegenden Reliabilitätsstudie nicht alle Aspekte, welche die Meßgenauigkeit von SAGE beeinflussen, erfaßt werden können, da bestimmte Probleme zum Beispiel erst im Zusammenhang mit Vergleichen unterschiedlicher Individuen an Einfluß gewinnen.

Dieser Abschnitt sollte deutlich machen, daß SAGE zahlreichen Einflüssen unterworfen ist, welche die Reliabilität dieser Methode negativ beeinflussen könnten. Kann den dargestellten Problemen, begegnet werden? Und wenn ja, wie? Diese Fragen sollen im Folgenden beantwortet werden.

5.2.4.6 Möglichkeiten zur Behandlung des zufälligen Fehlers

Exaktes Arbeiten unter möglichst gleichbleibenden Bedingungen und Modifikationen in der Praxis von SAGE, wie die im ersten Teil der Arbeit diskutierten, die zufällige Schwankungen minimieren, können die Reliabilität erhöhen. Ganz vermeiden lassen sich zufällige Fehler jedoch nicht - gerade bei einer komplexen Methode wie SAGE. Über diese allgemeinen Vorschläge hinaus ist der Zusammenhang von Meßgenauigkeit und Stichprobengröße (auf verschiedenen Ebenen) zu beachten.

Allgemein gilt: Bei geringer Reliabilität streuen Werte weiter als bei hoher. Das bedeutet, daß, je geringer die Reliabilität eines Verfahrens ausfällt, desto mehr Beobachtungen notwendig sind, um ein statistisch signifikantes Ergebnis im Rahmen eines üblichen SAGE Experimentes zu erzielen (Bortz et al. 1990, S. 60). Eine Lösungsmöglichkeit für zufällige Fehler, die im Laufe eines SAGE Durchlaufes entstehen, wäre also, die Stichprobengrößen auf verschiedenen Ebenen zu erhöhen. Was dies in bezug auf die verschiedenen Phasen von SAGE genau bedeutet, soll im folgenden dargestellt werden.

Im Kontext von SAGE können drei Arten von Stichproben unterschieden werden: die Anzahl der untersuchten Individuen, die Anzahl der parallel durchgeführten Durchläufe, die auf einem einzigen aufgeteilten RNS-Pool oder auf verschiedenen Pools basieren, und die Anzahl der sequenzierten Tags als Stichprobe der Gesamtpopulation "untersuchtes Transkriptom". Je

nachdem welche Phase oder Problematik betrachtet wird, ist eine andere Stichprobenart der Bezugpunkt.

Die der ersten Phase von SAGE, der Hauptphase aus der Sicht molekular-biologischer Praxis, zugrundeliegende Stichprobe ist die Anzahl der durchgeführten Durchläufe. Dies hieße, entweder eine (mehrfache) Spaltung der extrahierten RNS wie in der vorliegenden Arbeit durchzuführen oder diverse RNS-Pools zur Erstellung eines *einzigen* Expressionsprofils zu untersuchen.[71] Es stellt sich jedoch die Frage, inwiefern eine derartige Forderungen in einem Standardlabor umsetzbar ist.

Um die durch Polymorphismen des Transkriptoms und anderen interindividuellen Unterschieden induzierten zufälligen Schwankungen auszugleichen, müßte ein Expressionsprofil auf der Basis von vielen Individuen erstellt werden. Dies stellt - abgesehen von speziellen Varianten von SAGE, die kleine Mengen und damit einzelne Individuen untersuchen können - den Normalfall eines SAGE-Durchlaufes dar. Durch das Poolen werden die Schwankungen jedoch nicht quantifiziert. Ideal wäre auch hier, das Mitteln mehrer Profile von unterschiedlichen Individuen.

Die Streuung des zufälligen Fehlers bei der Sequenzierung der Polytagketten bezieht sich auf die Anzahl der sequenzierten Tags. Hier ermöglicht eventuell die Steigerung der Effizienz der Sequenzierautomaten einen Handlungsspielraum, um eine möglichst große Anzahl an Tags bei möglichst geringen finanziellen und personellem Aufwand zu erreichen.[72]

Um zu versuchen, diesen zufälligen Sequenzfehler für die vorliegenden Daten auf einem andere Weg zu reduzieren und so die Reliabilität a posteriori positiv zu beeinflussen, war ein Computerprogramm zur systematische Korrektur entwickelt worden, das Tags, welche nur einmal vorkamen, häufigeren Tags zuordnete, wenn ein Unterschied von einer Base bestand (siehe S. 73). Somit sollte die zufällige Streuung der Werte und die Verschiedenheit der Tagpaare reduziert werden. Im Folgenden wird die statistische Evaluation dieser Korrektur erörtert und die Korrektur bewertet werden.

---

[71] Diese Durchläufe sollten nicht miteinander addiert, sondern gemittelt werden!

[72] Auch die schon diskutierte Problematik der Stichprobenvariabilität, die in die Gesamtreliabilität zusätzlich zur methodischen mit einfließt, läßt sich über die Erhöhung dieser Stichprobe positiv beeinflussen.

Evaluation der Sequenzfehlerkorrektur

Im statistischen Vergleich der beiden Datensätze "mit" versus "ohne" diese Korrektur wurde geprüft, ob sich diese Korrektur in den Ergebnissen der paarweisen Signifikanztests insofern äußert, als in dem Datensatz mit Korrektur weniger Paare mit statistisch signifikanten Unterschieden zu finden seien. Dazu wurde folgendes Vorgehen gewählt. Es wurde die Anzahl der Tagpaare, die in Rahmen der Simulation üblicher SAGE Experimente einen statistisch bedeutsamen Unterschied aufweisen (simuliertes 5% Niveau), der beiden Datensätze einander gegenüber gestellt. Im Folgenden werden die Ergebnisse dieses Vergleichs diskutiert.

Es kann nicht nachgewiesen werden, daß durch die systematische Korrektur des Sequenzfehlers eine stärkere Homogenität zwischen K1 und K2 entsteht, wie aufgrund der theoretischen Überlegungen angenommen worden war. Über die Gründe hierfür lassen sich verschiedene Vermutungen anstellen, die kurz dargestellt werden sollen.

Wenn die Anzahl der Tagpaare verglichen wird, welche auf dem 5% Simulationsniveau statistisch verschieden erscheinen, dann ist - entgegen der aufgestellten Hypothese - in dem Datensatz *mit* Korrektur eine vergleichsweise höhere Anzahl dieser statistisch unterschiedlichen Paare finden. Das bedeutet, daß die in dieser Arbeit entwickelte Sequenzfehlerkorrektur nicht zur Erhöhung der Meßgenauigkeit beiträgt. Dies wirft die Frage auf, inwiefern die Korrektur unzureichend ist.

Da lediglich Tags zugeordnet werden, die eine Häufigkeit von 1 haben, wird nur ein Bruchteil des vorhandenen Fehlers ausgeglichen. Des weiteren erhalten Tags, die mindestens 3 mal vorkommen, gegebenenfalls durch die Korrektur einen höheren Wert. Fehlerhafte Tags dagegen, welche fälschlicherweise häufigen Tags zugerechnet werden, werden von diesen *nicht* abgezogen. Das heißt, daß die Korrektur assymetrisch erfolgt: von unten nach oben, nicht jedoch umgekehrt.

Ein weitere Erklärung ist folgende. Der Entwicklung der hier geprüften Hypothesen liegt die Annahme zugrunde, daß von einer Gleichheit der beiden Profile ausgegangen werden kann (globale H_0, siehe S. 113). Dann würde bei genauerem Messen durch die Korrektur eine Reduktion der Verschiedenheit zu erwarten sein. Wenn aber durch die Herstellung der SAGE Profile zufällige Ungenauigkeiten (vergleiche S. 166ff), welche die Reliabilität von SAGE senken, induziert werden, welche die beobachteten unterschiedlichen Taghäufigkeiten widerspiegeln, dann könnte das genauere Messen durch die Sequenzfehlerkorrektur diese Unterschiede deutlicher werden lassen. Das würde bedeuten, daß die hier vorgestellten Ergebnisse nicht gegen die Korrektur, sondern für sie sprechen würden. Diese Überlegung

reflektiert die Tatsache, daß aufgrund der Komplexität von SAGE zufällige Fehler, welche die Reliabilität der Methode beeinflussen, zu verschiedenen Zeitpunkten auftreten können und sich fortpflanzen.
Eine aufwendige Überprüfung der in der vorliegenden Arbiet entwickelten Fehlerkorrektur und der oben aufgestellten Erklärungsansätze wäre, SAGE Daten aus einfachen Sequenzierungen mit und ohne Korrektur mit Daten zu vergleichen, die aus doppelten Sequenzierungen stammen, so daß davon ausgegangen werden kann, daß nur ein minimaler Sequenzfehler vorliegt und diese Sequenzdaten als externe Kontrolle im Sinne einer Validierung fungieren können. Es könnte so überprüft werden, ob die a posteriori Korrektur eine Annäherung des Datensatzes an denjenigen des doppelt sequenzierten erbringt.
Als Fazit der Sequenzfehlerkorrektur ist folgendes zusammenzufassen: Solange die Ursachen der Ergebnisse des hier vorgestellten Vergleiches nicht bekannt sind, sollte die Korrektur in der in dieser Arbeit entwickelten Form nicht angewendet werden. Die Ergebnisse machen auch deutlich, welchen herausragenden Stellenwert das möglichst fehlerfreie Sequenzieren hat, da so sekundäre Korrekturen nicht notwendig werden.

5.2.4.7 Fazit der ermittelten Reliabilität

Die Reliabilität von SAGE, wie sie in der vorliegenden Arbeit ermittelt wurde, kann nicht als gut bezeichnet werden. Es bleibt jedoch offen, ob die ermittelten Unterschiede auf mangelnde methodische Meßgenauigkeit von SAGE oder aber auf stochastische Schwankungen, die der Stichprobenentnahme geschuldet sind, zurückzuführen sind.
Was bedeutet das für die Verwendung von SAGE zur Messung von Transkriptomen?
Es heißt nicht, daß SAGE nicht als wissenschaftliche Methode benutzt werden kann, sondern daß bestimmte Aspekte bei der Auswertung und Interpretation der Expressionsdaten berücksichtigt werden sollte. Diese werden im Folgenden zusammengefaßt.

- Erhöhung der Stichprobengrößen auf verschiedenen Ebenen:

Das Charakteristikum, daß SAGE ein Stichproben untersuchendes Verfahren ist, führt dazu, daß durch die Größe der Stichprobe und damit die Repräsentativität des konkreten Projektes die Meßgenauigkeit von SAGE bereits zu einem gewissen Grad determiniert ist. Je größer eine Stichprobe ist, umso repräsentativer wird das Projekt, und umso geringer fallen die stochastischen Schwankungen aus.
Die allgemeine Möglichkeit oder sogar Notwendigkeit, eine möglicherweise suboptimale methodische Reliabilität über Erhöhung von Stichprobengrößen auszugleichen, ist detailliert ausgeführt worden (siehe S.169). Es spielt dabei nicht nur die Gesamttagmenge eine Rolle,

sondern auch die Untersuchung mehrerer (geteilter) RNS-Pools bezeihungsweise Individuen, deren Resultate gemittelt (nicht addiert!) werden. Zu beachten ist bei der Erhöhung der Gesamttagmenge, daß so immer kleinere Differenzen signifikant sein können. Eine Planung des sogenannten optimalen Stichprobenumfangs mit festgelegtem alpha- und beta-Fehler wäre also sinnvoll (siehe S. 173).

- Vorrangige Analyse von häufig auftretenden Tags:

Je geringer das Expressionsniveau eines Transkriptes ist, desto größer fallen die stochastischen Schwankungen aus, die mit der Stichprobenentnahme einhergehen. Außerdem gilt: Größere Taghäufigkeiten tolerieren durch die Praxis von SAGE bedingte Meßungenauigkeiten besser, da bei gleichbleibender Größe des zufälligen Fehleranteils, der auf methodische Meßungenauigkeiten zurückzuführen ist, dieser relativ gesehen geringer ausfällt als bei Tags, die selten auftreten, so daß häufig auftretende Transkripte in der Praxis von SAGE besser zu handhaben sind. Dies spricht ebenfalls für die Sequenzierung einer möglichst großen Menge an Tags, da so die Häufigkeiten der einzelnen Transkripte ansteigen, und - wenn möglich - für die Auswertung von Tagpaaren, deren Mittelwert eine gewisse Mindesthäufigkeit hat. Letzteres ist in Projekten, die Gewebe und/oder Zustände mit komplexem Expressionsmuster untersuchen, unter Umständen nur schwer umzusetzen, da viele Gene niedrige Expressionsniveaus aufweisen. Dies bedeutet, daß SAGE vorzugsweise auf homogene Materialien (zum Beispiel Zellkultur) angewandt werden sollte.

- Nur Transkripte in die Auswertung einbeziehen, die einen ausgeprägten Regulationsunterschied aufweisen:

Wenn die Reliabilität von Messungen nicht optimal ist, können nur große Unterschiede reliabel gemessen werden. Bei kleinen Unterschieden, auch wenn diese statistisch signifikant werden, bleibt unklar, ob sie praktisch relevant sind und zuverlässig gemessen wurden. Dies bedeutet, daß ein - möglichst großer - Faktor festgelegt werden sollte, um den Transkripten mindestens reguliert sein sollten, um für weitere Untersuchungen relevant zu sein.

- Überwachung der Meßgenauigkeit in jedem SAGE Experiment:

Um zumindest auf qualitativer Ebene eine Idee der Genauigkeit jedes einzelnen SAGE Experimentes zu erhalten, wäre es zu empfehlen, den in dieser Arbeit vorgestellten Aufbau eines SAGE Durchlaufes, das heißt die parallele Verarbeitung eines mindestens einmal geteilten RNS Pools, als Standard zu etablieren.

Um die Schwankungen, die durch die stochastische Natur der Stichprobenentnahme entstehen, zu erfassen, wäre zusätzlich jeweils ein Vergleich der ersten Häfte der sequenzierten Tags mit der zweiten Hälfte erforderlich. Auf diese Weise würde eine Teilung

des Materials erst im letzten Schritt von SAGE erfolgen, so daß - unter der Voraussetzung, daß der Sequenzfehler minimiert worden ist - alle Meßungenauigkeiten induzierenden Schritte ausgeschlossen wären.

- Verwendung einer externen Kontrolle:

Wichtig ist zu beachten, daß angesichts der vorhandenen Unsicherheit bezüglich der Reliabilität von SAGE Resultate dieser Methode nicht für sich stehen sollten, sondern daß SAGE als sehr potentes Screeningverfahren aufgefaßt werden sollte, dessen ermittelte Expressionsunterschiede immer anhand eines zweiten etablierten Verfahren (beispielsweise Northern Blot) überprüft werden sollten.

5.2.4.8 Ausblick

Wenn SAGE solchermaßen als Screeningverfahren aufgefaßt wird, impliziert dies wie soeben erläutert, daß die Resultate dieser Methode steter Validierung mit einer zweiten Methode bedürfen. Hier bietet sich eine Möglichkeit für weitere Arbeiten zu SAGE: Die Überprüfung der Gesamtvalidität von SAGE steht noch aus.

In der vorliegenden Arbeit ist keine Aussage darüber möglich, ob die ermittelte Reliabilität deswegen nicht als gut eingestuft werden kann, weil SAGE nicht genau mißt, oder ob aufgrund der Gewebskomplexität und der relativ kleinen Stichprobe eine deutliche Stichprobenvariabilität vorliegt. Aus diesem Grund wären weitere Reliabilitätsstudien mit homogenen Geweben (zum Beispiel Zellkulturen) und größeren Stichproben zu empfehlen, die bei festgelegtem alpha- und beta-Fehler den sogenannten optimalen Stichprobenumfang jedoch nicht überschreiten. Es wäre interessant zu quantifizieren, welchen Anteil die stochastischen Schwankungen bei der Stichprobenentnahme haben und eine an einer noch zu tolerierenden Variabilität orientierte Mindeststichprobengröße festzulegen. Grundsätzlich wäre bei weiteren Reliabilitätsstudien die Anwendung eines Äquivalenztestes zu empfehlen.

Interessant wäre auch, anhand *mehrerer* parallel durchgeführter SAGE Durchläufe zu prüfen, wie groß die Fehlervarianz von SAGE ist (quantitativer Ansatz), und wie stark der Einfluss der Höhe des Expressionsniveaus ist. Auf diese Weise könnte auch untersucht werden, ob die zufälligen Fehler wirklich normalverteilt sind,[73] um so den geeigneten Kennwert zur Mittelung von Taghäufigkeiten ermitteln zu können.

[73] Nach Hart et al. (1997³) ist die Annahme, daß zufällige Fehler normalverteilt sind zwar oft gerechfertigt, es sei jedoch keine Naturgesetz.

6 Zusammenfassung

Etablierung

Die vorliegende Arbeit ist im Rahmen eines Projektes zur Untersuchung der Genexpression bei Tiermodellen neurologischer Erkrankungen entstanden. Mit herkömmlichen Kandidatenansätzen und den entsprechenden Methoden wie beispielsweise Northern Blotting ist eine Expressionsanalyse nur in beschränktem Umfang zu realisieren. Ziel war daher die Etablierung eines Verfahrens wie SAGE, das die Analyse des gesamten zerebralen Transkriptoms zuläßt. Wie die Arbeit gezeigt hat, ist SAGE in einem Standardlabor durchführbar. Das Verfahren hat viele Entwicklungsmöglichkeiten (beispielsweise LongSAGE, SAGE lite), so daß SAGE der Forschung auch weiterhin neue Impulse (zum Beispiel im Zusammenhang mit Einzelzellanalysen) geben kann.

Im Verlauf der Etablierung von SAGE wurden folgende Abwandlungen der Orginalmethode durchgeführt: Um wenige redundante Ditags, welche die Anzahl an auswertbaren Tags senken, zu erhalten, wurden im Rahmen der PCR erfolgreich parallel sehr viele Ansätze mit niedriger Zyklenzahl durchgeführt. Da sich die Ligation der Ditags zu Konkatemeren aufgrund der hohen Qualität der Ditags als übereffizient erwieß, genügte eine geringere Menge des Enzyms sowie wesentlich kürzere Inkubationszeiten. Die Begradigung der Tags nach dem BsmFI Verdau mit Klenow erfolgt bei niedrigerer Temperatur, um die Exonukleaseaktivität des Enzyms zu verringern und damit die Wahrscheinlichkeit für zu kurze Tags zu senken.

Es wurden verschiedene Problembereiche offensichtlich, zu welchen Vorschläge zur Lösung (auch aus der Literatur) diskutiert wurden. Hierzu zählen: die Minderung der Effizienz von SAGE durch Kontamination mit Linkersequenzen, der suboptimal verlaufenden Verdau mit NlaIII und Sonderfälle der Boten-RNS-Gestalt wie Transkripte ohne NlaIII-Schnittstelle oder einer sehr weit 5’ liegenden und Transkripte, deren Gegenstrang die BsmFI Erkennungssequenz enthält. Zur Problematik des CG-Gehalts der Tags, der in der vorliegenden Arbeit in den beiden Gruppen nicht übereinstimmt (statistisch abgesichert), sind weitere Studien notwendig.

Im Zusammenhang mit der Auswertung waren dies: Die Elimination der Linkerartefakte erschien nicht ausreichend, so daß hierfür ein spezielles Computerprogramm entwickelt wurde. Für Publikationen wäre zu wünschen, daß die Kriterien und die Art der Linkerelimination dokumentiert wäre. Die Homologierecherche kann sich schwierig

gestalten, da Uneindeutigkeiten der Zuordung auftreten können (Tags können mehreren Genen, ein Gen kann mehreren Tags, Tags können falsch oder gar nicht zugeordnet werden.). Lösungsansätze wurden dargestellt, insofern die Uneindeutigkeiten der Zuordnung nicht aufgrund der Datenlage der Genbanken auftreten, sondern SAGE anzulasten sind.

Statistik

Zur Evaluierung der statistischen Auswertung von SAGE wurde zusätzlich zu einer ausführlichen Darstellung des gesamten statistischen Entscheidungsprozesses explorativ die Situation statistischer Entscheidungen nachgeahmt, wie sie sich im Rahmen üblicher experimenteller Konstellationen darstellt. Es wurde eine Testvariante (modifizierter Z-Test) angewandt und evaluiert, die bis dato noch nicht zur Auswertung von SAGE benutzt worden war. Die Ergebnisse der Berechnungen zeigen die Eigenheiten der verschiedenen Tests. Dies macht deutlich, daß bei der statistischen Auswertung eines SAGE Projektes der Test gemäß den Eigenschaften des Projektes ausgewählt werden sollte. Es folgt eine entsprechende Zusammenfassung der vier evaluierten paarweisen Tests.

a) Test nach Madden et al. (1997)

Aufgrund der Tatsache, daß sich N_1 und N_2 für die gültige Anwendung dieses leicht zu berechnenden Tests identisch oder zumindest sehr ähnlich sein müssen, ist diese sehr eingeschränkt. Die Ergebnisse dieses Tests in der vorliegenden Arbeit unterscheiden sich nicht statistisch signifikant von denjenigen des Tests nach Audic und Claverie oder des Vier-Felder-Chi²-Testes. In einer weiteren Analyse erweist dieser Test sich als derjenige, von den in der vorliegenden Arbeit untersuchten Tests, welcher am konservativsten entscheidet.

b) Test nach Audic und Claverie (1997)

Im Gegensatz zu allen anderen vorgestellten Tests unterliegt der Test von Audic und Claverie keinerlei Einschränkungen bezüglich der Größe von N_1 und N_2 und der Taghäufigkeiten n. Diese Parameter können jeden beliebigen Wert annehmen. Im Bereich kleiner Taghäufigkeiten entscheidet der Test konservativ und hat eine geringe Teststärke, so daß dort eine niedrige Wahrscheinlichkeit vorliegt, daß der Test zugunsten von H_1 entscheidet. Diese Eigenschaften machen den Test von Audic und Claverie zu einem geeigneten Test, wenn keine minimale zu verwendende Taghäufigkeit festlegt werden soll, aber in diesem unteren Bereich, die Daten unter strengen Kriterien betrachtet und wenig falsch positive Transkripte riskiert werden sollen. Dies könnte zum Beispiel in SAGE Projekten mit einer vergleichsweise geringen Anzahl an sequenzierten Tags der Fall sein oder bei Projekten, die Gewebe oder Zustände mit einem komplexen Expressionsmuster untersuchen. Da keine

Software vorliegt, die eine rein statistische Analyse ganzer SAGE Projekte automatisiert durchführt, ist die Anwendung dieses Tests relativ aufwendig.

c) Vier-Felder-Chi²-Test

In der Monte-Carlo-Studie von Man et al. (2000) erweist sich dieser einfach zu berechnende Test als robust. Er entscheidet weder konservativ noch progressiv, sondern bleibt beim vorgegebenen α-Niveau. Seine Teststärke ist höher als diejenige des Tests nach Audic und Claverie. Dies gilt insbesondere für den Bereich kleiner Taghäufigkeiten. Die Daten der vorliegenden Arbeit bestätigen dies. Wenn die Tagpaare selektiert werden, deren Mittelwert $m < 15$ ist, ermittelt der Chi²-Test mehr Paare als statistisch signifikant verschieden als der Test nach Audic und Claverie. Im Bereich größerer Häufigkeiten entscheiden beide Tests identisch. Allerdings stellt sich im Bereich sehr kleiner Häufigkeiten die Frage nach der Verletzung der Voraussetzungen des Chi²-Testes, der Einschränkungen bezüglich der erforderlich minimalen Erwartungshäufigkeiten unterworfen ist. Dies macht den Test besonders geeignet für SAGE Projekte, die eine sehr große Anzahl von Tags sequenziert haben und/oder ein Gewebe oder einen Zustand untersuchen, das/der ein einfaches Expressionsmuster besitzt, so daß auf die Auswertung sehr kleiner Taghäufigkeiten verzichtet werden kann.

d) Z-Test

Die statistischen Entscheidungen des von Kal et al. (1999) vorgestellten Ansatzes, der auch der Software SAGEstat zugrunde liegt, stimmen exakt mit denjenigen des Vier-Felder-Chi²-Test überein. Da die Anwendung dieses Z-Testes ebenfalls Einschränkungen unterliegt, was die minimale zu verwendende Häufigkeit n betrifft, ist die Anwendung des modifizierten Z-Testes zu empfehlen. Bei dieser Variante des Z-Testes ist nur die Häufigkeit $n = 0$ problematisch. Der Vergleich der untersuchten vier Tests macht deutlich, daß die Ergebnisse des modifizierten Z-Test deutlich von denjenigen der anderen Tests insbesondere im Bereich kleiner Taghäufigkeiten abweichen. Er entscheidet hier deutlich progressiver. Wenn angestrebt wird, mittels SAGE möglichst viele Gene als reguliert zu identifizieren, die weiteruntersucht werden sollen, ist dieser Test damit geeignet. Dies gilt auch für große SAGE Projekte, die auf die Auswertung von Tags mit geringen Häufigkeiten (insbesondere $n = 0$) verzichten können.

Reliabilität

Um die Reliabilität von SAGE abschätzen zu können, wurde von vier Mäusegroßhirnen die Gesamt-RNS extrahiert und vereinigt. Diese Transkriptgrundpopulation wurde zweigeteilt

und parallel untersucht. Die beiden Gruppen wurden anhand eines statistischen Tests, der die gesamte Verteilungen der beiden Profile prüft, auf Homogenität untersucht. Zusätzlich wurde der Zusammenhang (und dessen Ausmaß) der Profile ermittelt. Dies ergab, daß die Reliabilität von SAGE im vorliegenden Kontext (relativ geringe Stichprobe und ein komplexes Gewebe) nicht optimal ist. Es kann jedoch keine Aussage dazu gemacht werden, ob dies der Methode selbst, das heißt ihrer molekularbiologischen Praxis und der Datenaufbereitung, anzulasten ist oder einer großen Stichprobenvariabilität. Dies bedeutet, daß in der vorliegenden Arbeit keine endgültige Aussage zur Reliabilität von SAGE möglich ist.

Es gibt Möglichkeiten eine eventuell suboptimale Reliabilität im Rahmen von zukünftigen Projekten auszugleichen, indem die Stichprobengröße (Anzahl der sequenzierten Tags und die Anzahl der Durchläufe, deren Ergebnis *gemittelt* wird) erhöht wird. Außerdem ist zu beachten, daß dies bedeutet, daß nur große Regulationsunterschiede und Transkripte höherer Expressionsniveaus (die bei der Untersuchung homogener Gewebearten häufiger sind) so gemessen werden können, daß die Rate der falsch positiven Entscheidungen im Rahmen des vorgegebenen α-Signifikanzniveaus bleibt. Um im Rahmen eines SAGE Projektes die Reliabilität des Durchlaufs zumindest qualitativ beurteilen zu können, wäre es zu empfehlen, - wie in der vorliegenden Arbeit - die verwendete Gesamt-RNS zu teilen, parallel zu verarbeiten und die beiden Profile auf Homogenität zu prüfen. Aus dem gleichen Grund wäre es sinnvoll, ein Verfahren zur Abschätzung des Sequenzfehlers, der die Reliabilität von SAGE so wesentlich beeinflußt, standardmäßig anzuwenden.

Die in der vorliegenden Arbeit entwickelte Sequenzfehlerkorrektur zur Verbesserung der Meßgenauigkeit zeigte keine Erhöhung der Homogenität der beiden Vergleichsprofile K1 und K2. Solange noch keine exakte Evaluation dieser Korrektur in einem speziellen Versuch erfolgt ist oder eine andere entwickelt worden ist, ist keine nachträgliche Korrekturmöglichkeit vorhanden, was die Forderung nach hochwertiger Sequenzierung der SAGE Konkatemere betont.

Aufgrund der genannten Einschränkungen der Reliabilität sollte SAGE als potentes Screeningverfahren im Sinne einer explorativen Datenanalyse aufgefaßt werden, dessen Resultate steter Validierung mit einer zweiten Methode bedürfen.

Literaturverzeichnis

Adams, M.D., Kelley, J.M. et al. (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. Science *252*, 1651 - 1656.

Adams, M.D., Kerlavage, A.R. et al. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature [Suppl] *377*, 3 - 173.

Altman, D.G. (1991). Practical statistics for medical research. London.

Angelastro, J.M., Klimaschewski, L. et al. (2000a). Identification of diverse nerve growth - regulated genes by serial analysis of gene expression (SAGE) profiling. PNAS *97* (19), 10424 - 10429.

Angelastro J. M., Klimaschewski L.P., Vitolo O. V. (2000b). Improved NlaIII digestion of PAGE-purified 102 bp ditags by addition of a single purification step in both the SAGE and microSAGE protocols. Nucleic Acids Res. *28* (12), E62.

Arkin, A., Ross, J., McAdams, H.H. (1998). Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage Lambda - infected Escherichia coli Cells. Genetics *149*, 1633 - 1648.

Audic, S. und Claverie, J.-M. (1997). The Significance of Digital Gene Expression Profiles. Genome Research *7*, 986 - 995.

Baas, F., Tabak, H. (1999). A tale of tags: report on a HUGO/EU SAGE Workshop, 29 January - 1 February 1999, Hilversum, The Netherlands. Eur J Hum Gen *7*, 510 - 512.

Bachem, C.W., van der Hoeven, R.S. et al. (1996). Visulization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. Plnt J *9* (5), 745 - 753.

Backert, S., Gelos, M. et al. (1999). Differential gene expression in colon carcinoma cells and tissues detected with a cDNA array. Int J Canver *82* (6), 868 - 874.

Basson, M.D., Liu, Y.W. et al. (2000). Identification and comparative analysis of human colonocyte short - chain fatty acid response genes. J Gastrointest Surg *4* (5), 501 - 512.

Benjamini, Y. und Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc *57* (1), 289 – 300.

Bernard, K., Auphan, N. et al. (1996). Multiplex messenger assay: simultaneous, quantitative measurement of expression of many genes in the context of T cell activation. Nucleic Acid Research *24* (8), 1435 - 1442.

Bertelsen, A. und Velculescu, V. (1998): High-throughput gene expression analysis using SAGE. Drug Discovery Today *3* (4), 152-159.

Bertucci, F., Van Hulst, S. et al (1999a). Expression scanning of an array of growth control genes in human tumor cell lines. Oncogene *18* (26), 3905 - 3912.

Bertucci, F. Bernard, K. et al. (1999b). Sensitivity issues in DNA array based expression measurements and performance of nylon microarrays for small samples. Hum Mol Genet *8* (9), 1715 - 1722.

Boccaccio, G. L., Carminatti, H. und Colman, D. R. (1999). Subcellular fractionation and association with the

cytoskeleton of messengers encoding myelin proteins. J Neurosci Res *58* (4), 480 - 491.

Bortz, J. und Döring, N. (1995²). Forschungsmethoden und Evaluation. Berlin, Heidelberg, New York.

Bortz, J. (1993[4]). Statistik für Sozialwissenschaftler. Berlin.

Bortz, J., Lienert, G. und Boehnke, K. (1990). Verteilungsfreie Methoden in der Biostatistik. Berlin, Heidelberg, New York.

Brady, G. (2000). Expression profiling of single mammalian cells - small is beautiful. Yeast *17* (3), 211 - 217.

Brenner, S., Johnson, M. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbeads arrays. Nat Biotechnol *28* (86), 630 - 634.

Carulli, J., Artinger, M. et al. (1998). High throughput Analysis of Differential Gene expression. J Cell Biochem Supple *30/31*, 286 - 296.

Chee, M., Yang, R. et al. (1996). Accessing genetic information with high-density DNA arrays. Science *274* (5287), 610 - 614.

Chen, H., Centola, M. et al. (1998). Characterization of gene expression in resting and Activated Mast Cells. J Exp Med *188* (9), 1657 - 1668.

Chen, J-J., Rowley, J.D., Ming Wang, S. (1999). Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. PNAS *97* (1), 349 - 353.

Chen, J-J., Sun, M. et al. (2002). Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. PNAS *99* (19), 12257 - 12262.

Cho, R.J., Fromont-Racine, M. et al. (1998). Parallel analysis of genetic selections using whole genome oligonucleotide arrays. PNAS *95*, 3752 - 3757.

Chrast, R., Scott, H.S. et al. (2000). The Mouse Brain Transcriptome by SAGE: Differences in Gene Expression between P30 Brains of the Partial Trisomy 16 Mouse Model of Down Syndrome (Ts65Dn) and Normals. Genome Res *10*, 2006 - 2021.

Cox, J.M. (2001). Applications of nylon membrane arrays to gene expression analysis. J Immunol Methods *250* (1-2), 3 - 13.

Croix, B., Velculescu, V. et al. (2000). MikroSAGE. Detailed Protocol. *Version 1.0e*

Daly, L.E. und Bourke, G.J. (2000[5]). Interpretation and uses of medical statistics. Oxford, Berlin, Tokyo.

Datson, N., van der Perk-de Jong, J. et al. (1999). MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acid Research *27*, 1300 - 1307.

Dawber, T.R. (1980). The Framingham Study. Cambridge (USA).

de Ferra, F., Engh, H. et al. (1985). Alternative splicing accounts for the four forms of myelin basic protein. Cell *43* (3 Pt 2), 721 - 727.

Dokas, L.A. (1983). Analysis of brain and pituary RNA metabolism: a review of recent methodologies. Brain Res *286* (2), 177 - 218.

Divina, P. und Forejt, J. (2004). The mouse SAGE site: database of public mouse SAGE libraries. Nucleic Acids

Research *32*, D 482 – 483.

Dudoit, S. et al. (2003). Mulitple hypothesis testing in microarray experiments. Statistical Science *18* (1), 71 – 103.

Eickhoff, H., Schuchhardt, J. et al. (2000). Tissue gene expression analysis using arrayed normalized cDNA libraries. Genome Res *10* (8), 1230 - 1240.

Ewing, B., Hillier, L. et al. (1998). Base calling of automated sequencers traces using phred. Accuracy assesment. Genome Res. *8* (3), 175 - 185.

Fields, S., Kohara, Y., Lockhart, D.J. (1999). Functional genomics. PNAS *96*, 8825 - 8826.

Fisher, L. und van Belle, G. (1993). Biostatistics: a methodology for the health sciences. New York.

Fodor, S.P., Read, J.L. et al. (1991). Light-directed, spatially addressable parallel chemical synthesis. Science *251* (4995), 767 - 773.

Geng, M.M., Ellenrieder, V. et al. (1999). Use of representational differncs analysis to study the effect of TGFB onthe expression profile of a pancreatic cancer cell line. Genes Chromosomes Cancer *26* (1), 70 - 79.

Geng, M.M., Wallrapp, C. et al. (1998). Isolation of differentially expressed genes by combining representational difference analysis (RDA) and cDNA library arrays. Biotechniques *25* (3), 434 - 438.

Green, C.D., Simons J.F. et al. (2001). Open systems: panoramic views of gene expression. J Immunol Methods *250* (1-2), 67 - 79.

Gress, T.M., Hoheisel, J.D. et al. (1992). Hybridisation fingerprinting of high-density cDNA-library arrays with cDNA pools from whole tissues. Mamm Genome *3* (11), 609 - 619.

Gress, T.M., Wallrapp, C. et al. (1997). Identifikation of genes with specific expression in pancreatic cancer by cDNA representional difference analysis. Genes Chromosomes Cancer *19* (2), 97 - 103.

Gygi, S., Rochon, Y. et al. (1999). Correlation between Protein and mRNA Abundance in Yeast. Mol Cell Biol *19* (3), 1720 - 1730.

Hacia, J.G., Woski, S.A. et al. (1998). Enhanced high density oligonucleotide array-based sequence analysis using modifies nucleoside triphosphates. Nucleic Acid Res *26* (21), 4975 - 4982.

Hastie, N.D. und Bishop, J.O. (1976). The expression of three abundance classes of messenger RNA in mouse tissues. Cell *9* (4 PT 2), 761-774.

Hart, H. (1997[7]). Meßgenauigkeit. München, Wien, Oldenburg.

Hartung, J. (1995[10]). Statistik: Lehr- und Handbuch der angewandten Statistik. Oldenburg.

Hashimoto, S.-I., Takuji, S. et al. (1999). Serial Analysis of Gene Expression in Human Monocytes and Macrophages. Blood *94* (3), 837 - 844.

Hastie, N.D. und Bishop, J.O. (1976). The expression of three abundance classes of messenger RNA in mouse tissues. Cell *9*, 761 - 774.

Hauser, N.C., Vingron, M. et al. (1998). Transcriptional profiling on all open reading frames of Saccharomyces cervesiae. Yeast *14* (13), 1209 - 1221.

Heller, R.A., Schena, M. et al. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. PNAS *94*, 2150 - 2155.

Hieter, P. and Boguski, M. (1997). Functional Genomics: It´s All How You Read It. Science *278*, 601 - 602.

Hogg, R.V. und Craig, A.T. (1995). Introduction to Mathematical Statistics. New Jersey.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scand J Statist *6*, 65 – 70.

Hosack, D. et al. (2003). Identifying biological themes within lists of genes with EASE. Genome Biology *4*, R70.

Hauser, M. et al. (2003). Genomic convergence: identifying candidate genes for Parkinson`s disease by combining serial analysis of gene expression and genetic linkage. Human Molecular Genetics *12* (6), 671 – 676.

Hough, C.D., Sherman-Baust, C.A. et al. (2000). Large-scale serial analysis of gene expression reveals genes differentially expressed in ovarian cancer. Cancer Res *60* (20), 6281 - 6287.

Hubank, M. und Schatz, D.G. (1999). cDNA representional difference analysis: a sensitiv and flexible method for identification of differentially expressed genes. Methods Enzymol *303*, 325 - 349.

Hubank,, M. und Schatz, D.G. (1994). Identifying differences in mRNA expression by representional difference analysis of cDNA. Nucl Acid Res *22* (25), 5640 - 5648.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the humane genome. Nature *409*, 860 - 921.

Invanov, I., Schaab, C. et al. (2000). DNA micrarray technology and and antimicrobial drug discovery. Pharmacogenomics *1* (2), 169 - 178.

Ishii, M., Hashimoto, S. et al. (2000). Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. Genomics *68*, 136 - 143.

Jakob, A.N., Kalapurakal, J. et al. (1999). A receptor tyrosine kinase, UFO/Axl, and other genes isolated by a modified differential disply PCR are overexpressed in a metastatic prostatic carcinoma cell line DU 145. Cancer Detect Prev *23* (4), 325 - 332.

Kahn, J. und Mehraban, F. (2000). Gene exprssion profiling in an in vitro model of angiogenesis. Am J Pathol *156* (6), 1887 - 1900.

Kal A.J., van Zonneveld A.J., et al. (1999). Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Mol Biol Cell *10* (6), 1859-72.

Kenzelmann, M., Muhlemann, K. (1999). Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol. Nucleic Acid Research *27* (3), 917 - 918.

Kierzek, A., Zaim, J. und Zielenkiewicz, P. (2001). The effect of transcription and translation initation frequencies on the stochastic fluctuations in prokaryotic gene epression. J Biol Chem, *276* (11), 8165-8172.

Ko, M.S. (1992). Induction mechanism of a single gene molecule: stochastic or deterministic ? Bioessays *14* (5), 341 - 346.

Kozian, D., Kirschbaum, B. (1999). Comparative Gene Expression Analysis. Trends Biotechnol *17* (2):, 73 - 78.

Krause, B. und Metzler, P. (1984). Statistik. Berlin.

Lal, A., Lash, A.E., et al. (1999). A public database for gene expression in human cancers. Cancer Res *59* (21), 5403-5407.

Larsson, M., Stahl, S., et al. (2000). Expression profile Viewer (ExProView): A Software Tool for Transcriptome Analysis. Genomics *63*, 341 - 353.

Lash, A.E., Tolstoshev, C.M. et al. (2000). SAGEmap: A Public Gene Expression Resource. Genome Res *10*, 1051 - 1060.

Lavery, D.J., Lopez-Molina, L. et al. (1997). Selective amplification via biotin- and restristion mediated enrichment (SABRE), a novel selective amplification procedure for detection of differetially expressed mRNA. PNAS *94* (13), 6831 - 6836.

Lee, S., Clark, T., et al. (2002). Correct identification of genes from serial analysis of gene expression tag sequences. Genomics *79* (4), 598-602.

Liang, P. und Pardee, A.B. (1992). Differential display of eucaryotic messenger RNA by means of the polymerase chain reaction. Science *257*, 967 - 970.

Lienert, G. und Raatz, U. (1994[5]). Testaufbau und Testanalyse. München.

Lipshutz, R.J., Fodor, S.P., et al. (1999). High density synthetic oligonucleotide arrays. Nat genet *21* (1), 20 - 24.

Lockhart, D.J., Winzler, E.A. (2000). Genomics, gene expression and DNA arrays. Nature *405* (6788), 827 - 836.

Lorenz, R. (1996[4]). Grundbegriffe der Biometrie. Stuttgart, Jena, Lübeck, Ulm.

Lund, A.H., Duch, M. et al. (1996). Increased cloning efficiency by temperature-cycle ligation. Nucl Acids Res *24* (4), 800 - 801.

Madden SL, Galella EA, et al. (1997). SAGE transcript profiles for p53-dependent growth regulation. Oncogene *15*, 1079-1085.

Madden SL, Wang CJ, Landes G.. (2000). Serial analysis of gene expression: from gene discovery to target identification. Drug Discov Today. *5* (9), 415-425.

Man, M.Z., Wang, X. und Wang, Y. (2000). POWER_SAGE: comparing statistical test for SAGE experiments. Bioinformatics *16* (11), 953 - 959.

Margulies, E., Innis, J. (2000). eSAGE: managing and analysing data generated with Serial Analysis of Gene Expression (SAGE). Bioinformatics *16*, 650 - 651.

Margulies, E., Kardia, S. und Innis, J. (2001). Identification and prevention of a GC content bias in SAGE libraries. Nucleic Acid Res *29* (12), e60.

Martin, K.J. und Pardee, A.B. (2000). Identifying expressed genes. PNAS *97* (8), 3789 - 3791.

Matsumura, H. Nirasawa, S., Terauchi, R. (1999). Technical advance: transcript profiling in rice (Oryza sativa

L.) seedlings using serial analysis of gene expression. Plant J. *20*, 719 - 726.

McAdams, H.H. und Arkin, A. (1997). Stochastic mechanisms in gene expression. PNAS *94*, 814 - 819.

Meier-Ewert, S., Lange, J. et al (1998). Comparative gene expression profiling by oligonucleotide fingerprinting. Nucleic Acids Res *26* (9), 2216 - 2223.

Mercatante, D. R. et al. (2001). Modification of alternative splicing by antisense oligonucleotides as a potential chemotherapy for cancer and other diseases. Curr Cance Drug Targets *1* (3), 211 - 230.

Michiels, E.M., Oussoren, E., et al. (1999). Genes differentially expressed in medullablastoma and fetal brain. Physiol Genomics *1* (2), 83 - 91.

Money, T., Reader, S. et al. (1996). AFLP-based mRNA fingerprinting. Nucl Acids Res *24* (13), 2616 - 2617.

Munasinghe, A., Patankar, S. et al. (2001). Serial analysis of gene expression (SAGE) in Plasmodium falciparum: application of the technique to A-.T rich genomes. Mol Biochem Parasitol *113* (1), 23 - 34.

Nacht M, Ferguson AT, et al. (1999). Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. Cancer Res *59* (21), 5464-70.

Neilson, L., Andalibi, A. et al. (2000). Molecular phenotype of the human oocyte by PCR-SAGE. Genomics *63*(1), 13 - 24.

Newlands, S., Levitt, L. et al. (1998). Transcription occurs in pulses in muscle fibers. Genes and Developement *12*, 2748 - 2758.

Nguyen, C., Rocha, D. et al (1995). Differential gene expression in the murine thymua assayed by quantitative hybridisation of arrayed cDNA clones. Genomics *29* (1), 207 - 216.

Patanjali, S.R., Parimoo, S., Weissman, S.M. (1991). Construction of a uniform-abundance (normalized) cDNA library. PNAS *88* (11), 1943 - 1947.

Patino, W., Mian, O., Hwang, P. (2002). Serial analysis of gene expression: technical considerations and applications to cardiovascular biology. Circ Res *91* (7), 565 - 569

Pease, A., Solas, D. et al. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. PNAS *91* (11), 5022 - 5026.

Peters, D., Kassam, A. et al. (1999). Comprehensive transcript analysis in small quantities of mRNA by SAGE-Lite. Nucleic Acid Research *27* (24), 39.

Piquemal, D., Commes, T. et al. (2002). Transcriptome analysis of monocytic leukemia cell differentiation. Genomics *80* (3), 361 - 371.

Polyak, K., Xia, Y., et al. (1997). A model for p53-induced apaptosis. Nature *389*, 300 - 305.

Powell, J. (1998). Enhanced concatamer cloning: a modification to the SAGE /Serial Analysis if Gene Expression) technique. Nucleic Acid Res *26*, 3445 - 3446.

Prashar, Y. und Weissman, S.M. (1996). Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. PNAS *93* (2), 659 - 663.

Prashar, Y. und Weissman, S.M. (1999). READS: a method for display of 3'-end fragments of restriction enzyme-digestion of differetial gene expression. Methods Enzymol *303*, 258 - 272.

Rampon, C., Jiang, C.H. et al. (2000). Effects of eviromental enrichment on gene expression in the brain. PNAS *97* (23), 12880 - 12884.

Rasch, D. (Hrsg.) (1996). Verfahrensbibliothek. Bd. 1. Versuchsplanung und -auswertung. München, Wien, Oldenburg.

Ross, I., Broane, C., Hume, D. (1994). Transcription of individual genes in eukaryotic cells occurs randomly and infrequently. Immunol Cell Biol *72* (2), 177 - 185.

Ruijter, J.M. (1999). SAGE and Statistics. SAGE Workshop, 29 January - 1 February 1999, Hilversum, The Netherlands.

Ryo, A., Kondoh, N. et al. (2000). A Modified Serial Analysis of Gene Expression That Generates Longer Sequence Tags by Nonpalindromic Cohesiv Linker Ligation. Analytical Biochemistry *277*, 160 - 162.

Sachs, L. (1999[9]). Angewandte Statistik: Anwendung statistischer Methoden. Berlin, Heidelberg, New York, Tokyo.

Saha, S., Sparks, A.B., et al. (2002). Using the transcriptome to annotate the genome. Nat Biotechnol *20* (5), 508-12.

Sambrook, J., Fritsch, E.F., und Maniatis, T. (1989[2]). Molecular Cloning: a labatory manual. Cold Spr Harb, NY.

Sandberg, R., Yasuda, R. et al. (2000). Regional and strain-specific gene espression mapping in the adult mouse brain. PNAS *97* (3), 11038 - 11043.

Schena, M., Heller, R.A. et al. (1998). Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol *16* (7), 301 - 306.

Schena, M., Shalon, D. et al. (1996). Parallel human genome analysis: Mikroarray-based expression monitoring of 1000 genes. PNAS *93*, 10614 - 10619.

Schena, M., Shalon, D., et al. (1995). Quantitative monitoring of gene expression pattrens with a complementary DNA microarray. Science *270* (5235), 467 - 470.

Selinger, D.W., Cheung, K.J. et al. (2000). RNA expression analysis using a 30 base pair resolution E. coli genome array. Nat Biotechnol *18* (12), 1262 - 1268.

Shimkets, R.A., Lowe, D.G. et al. (1999). Gene expression analysis by transcript profiling coupeled toa gene database query. Nat Biotechnol *17* (8), 798 - 803.

Shimkets, R.A., Lowe, D.G. (1999). Gene expression analysis by transkript profiling couled to a gene database query. Nat Biotechnol *17* (8), 798 - 803.

Shklyaev, S., Namba, H. et al. (2000). SAGE transcript profiles in cultured human fetal fibroblasts, WI-38. DNA Seq *11* (3-4), 281 - 286.

Siegel, S. (1956). Nonparametric methods for the behavioral sciences. New York.

Soares, M.B., Bonaldo, M.F. (1994). Construction and characterization of a normalized cDNA library. PNAS *91* (20), 9228 - 9232.

Southern, E.M., Maskos, U., Elder, J.K. (1992). Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimantal models. Genomics *13* (4),

1008 - 1017.

Spanakis, E. (1993). Problems related to the interpretation of autoradiographic data on gene expression using common constitutiv trnascripts as controls. Nucleic Acid Research *21* (16), 3809 - 3819.

Spanakis, E. und Bouty - Boyé, D. (1994). Evaluation of quantitativ variation in gene expression. Nucleic Acid Research *22* (5), 799 - 806.

Spiegelman, D. (1998). Reliability Study, in: Armitage, P. und Colton, Th. (Hrsg). Encyclopedia of Biostatistics. London.

Spinella, D.G., Bernardino, A.K. et al. (1999). Tandem arrayed ligation of expressed sequence tags (TALEST): a new method for generating global gene expression profiles. Nucl Acid Res *27* (18), e22

Storey, J.D. (2002). A direct approach to false discovery rates. J R Statistical Soc *64*, 474 – 498.

Sutcliffe, J.G. (1988). mRNA in the mammilian central nervous system. Annu Rev Neurosci *11*, 157 - 198.

Sutcliffe, J.G., Foye, P.E. et al. (2000). TOGA: An automated parsing technology for analysing expression of nearly allgenes. PNAS *97* (5), 1976 - 1981.

Thibault, C., ChaoQiang, L. et al. (2000). Expression Profiling of Neuronal Cells Reveals Specific Patterns of Ethanol-responsive Gene Expression. Mol Pharmacol *52* (6), 1593 - 1600.

Thomas, E.A., Danielson, P.E. et al. (2001). Clozapine increases apolipoprotein D expression in rodent brain: towards a mechanism for neuroleptic pharmacotherapy. J Neurochem *76* (3), 789 - 796.

Trendelenburg, G. (1998). Expressionsuntersuchung einer Kolonkarzinom-Genbibliothek und Identifizierung eines neuen A-Kinase-Anker-Proteins (AKAP149) mit einer RNA-Bindungs-Domäne. Inaugural-Dissertation, Freie Universität, Berlin.

Tyagi, S. (2000). Taking a census of mRNA populations with microbeads. Nature Biotechnol *28* (86), 597 – 598.

van der Berg, A., Van der Leij., D., Poppema, S. (1999). Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. Nucleic Acid Res *27*, e17.

van Limpt, V., Chan, A. et al. (2000). SAGE Analysis of Neuroblastoma Reveals a High Expression of the Human Homologue of the Drosophila Delta Gene. Medical and Pediatric Oncology *53*, 554 - 558.

Velculescu V.E., Zhang L, et al. (1995). Serial Analysis Of Gene Expression. Science *270*, 484-487.

Velculescu, V.E., Zhang L., et al. (1997a). Characterization of the yeast transcriptome. Cell *88*, 243-251.

Velculescu, V.E., Zhang, L. et al. (1997b). Serial Analysis of Gene Expression. Detailed Protocol. Version 1.0c. Erhältlich durch: John Hopkins Oncology Center and Howard Hughes Medical Institute, 424 North Bond St, Baltimore, MD 21231, USA.

Velculescu, V.E. (1999a). Tantalizing Transcriptomes - SAGE and Ist Use in Global Gene Expression. Science *286*, 1491 - 1492.

Velculescu, V.E., Madden S. et al. (1999b). Analysis of human transcriptomes. Nature Genetics (23), 387 - 388.

Velculescu, V.E., Zhang, L. et al. (2000). Serial Analysis of Gene Expression. Detailed Protocol. Version 1.0e. Erhältlich durch: John Hopkins Oncology Center and Howard Hughes Medical Institute, 424 North Bond St, Baltimore, MD 21231, USA.

Vingron, M. und Hoheisel, J. (1999). Computational aspects of expresssion data. J Mol Med *77* (1), 3-7.

Virlon, B., Cheval, L. et al. (1999). Serial microanalysis of renal transcriptomes. PNAS *96* (26), 15286 - 15291.

Wallrapp, C., Müller-Pillasch, F. (1999). Novel technology for detection of genomic and transcriptional alterations in pancreatic cancer. Ann Oncol *10* (4), 64 - 68.

Wang, D.G., Fan, J.-B. et al. (1998). Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. Science *280*, 1077 - 1082.

Wang, S.M., Rowley, J.D. (1998). A strategy for genome-wide gene analysis: Integrated procedure for gene identification. PNAS *95* (29), 11909 - 11914.

Weineck, J. (1996[9]). Optimales Training: Leistungsphysiologische Trainingslehre unter besonderer Berücksichtigung des Kinder- und Jugendalters. Balingen.

Welle, S., Bhatt, K., Thornton, C.A. (1999). Inventory of high-abundance mRNAs in skeletal muscle of normal men. Genome Res. *9* (5), 506-13.

Welle, S., Bhatt, K., Thornton, C.A. (2000). High-abundance mRNAs in human muscle: comparison between young and old. J Appl Physiol. *89* (1), 297-304.

Wellek, S. (1994). Statistische Methoden zum Nachweis von Äquivalenz. Stuttgart.

Wellek, S. (2003). Testing Statistical Hypotheses of Equivalence. Boca Raton.

Welsh, J.B., Zarrinkar, P.P. et al. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. PNAS *98* (3), 1176 - 1181.

Westfall, P.H. und Young, S.S. (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. New York.

Wilcox, A.S., Khan, A.S., et al. (1991). Use of 3´untranslated sequences of human cDNA for rapid chromosome assignment and conversion to STSs: Implications for an expression map of the genome. Nucleic Acid Res. *19*, 1837 - 1843.

Wodicka, L., Dong, H. et al (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol *13*, 1359 - 1367.

Yang, Y. und Speed, T. (2003). Design and analysis of comparative microarray experiments. In: Speed, T. (Hrsg.). Statistical analysis of gene expression microarray data. Boca Raton.

Yamamoto, M., Wakatsuki, T. et al. (2001). Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods *250* (1-2), 45 - 66.

Ye, S.Q. und Zhang, L.Q. (2000). MiniSAGE: Gene expression profile using serial analysis of gene expression from 1 µg total RNA. Anal Biochem *287* (1), 144 - 152.

Yu, J., Zhang, L. et al. (1999). Identification and classification of p53-regulates genes. PNAS *96* (25), 14517 - 14522.

Zhang, L., Zhou, W., et al. (1997). Gene Expression Profiles in Normal and Cancer Cells. Science *276*, 1268-1272.

Danksagung

An dieser Stelle möchte ich allen, die an dem Zustandekommen dieser Arbeit beteiligt waren, danken.
Herrn Prof. Dr. U. Dirnagl danke ich für die Vergabe der Themas.
Besonderer Dank gilt Dr. G. Trendelenburg für die Einweisung in molekular-biologische Techniken und Arbeitsweisen, für die eingehende Beratung sowie für aufschlußreiche Hinweise und Kommentare. Ohne sein Entgegenkommen und seine Hilfe wäre diese Arbeit nicht durchführbar gewesen.
Dr. Malte Mienert und Dr. Oliver Redner ist für die konstruktiven Diskussionen vieler statistischer Detailprobleme zu danken, Dr. Oliver Redner außerdem für die Kooperation bei der Erstellung der Programme zur Datensortierung. Ihre Kritik eröffnete mir immer wieder neue Perspektiven.
Mein herzlicher Dank gilt allen MitarbeiterInnen der Abteilung für Experimentelle Neurologie der Charité für freundliche theoretische und praktische Hilfe im Laboralltag.
Für das Korrekturlesen danke ich Johanna Castell, Dr. Rolf Castell, Dr. Dirlich, Dr. Lars Morawietz, Benjamin Meyer-Krahmer, Stefanie Kessner und Loretta Ihme.
Schließlich danke ich meinen FreundInnen für ihr Verständnis und ihre Unterstützung, vor allem in schwierigen Phasen der Dissertation.

Lebenslauf

12.8.1973	Geburt in München
1980 - 1993	Schule
1993 - 2000	Studium der Humanmedizin
1997 - 2006	Dissertation
Seit 2003	ÄiP und Assistenzärztin HUK, Sommerfeld

Eidestattliche Erklärung

Die vorliegende Dissertation wurde von mir selbst ohne die (unzulässige) Hilfe Dritter verfaßt. Sie stellt auch in Teilen keine Kopie anderer Arbeiten dar. Die benutzten Hilfsmittel sowie die Literatur sind vollständig angegeben.